

Likelihood Component Analysis

Benjamin B. Risk^{1,2}, David S. Matteson¹, David Ruppert¹

¹Department of Statistical Science, Cornell University

²Statistical and Applied Mathematical Sciences Institute

November 6, 2015

Abstract

Independent component analysis (ICA) is popular in many applications, including cognitive neuroscience and signal processing. Due to computational constraints, principal component analysis is used for dimension reduction prior to ICA (PCA+ICA), which could remove important information. The problem is that interesting independent components (ICs) could be mixed in several principal components that are discarded and then these ICs cannot be recovered. To address this issue, we propose likelihood component analysis (LCA), a novel methodology in which dimension reduction and latent variable estimation is achieved simultaneously by maximizing a likelihood with Gaussian and non-Gaussian components. We present a parametric LCA model using the logistic density and a semi-parametric LCA model using tilted Gaussians with cubic B-splines. We implement an algorithm scalable to datasets common in applications (e.g., hundreds of thousands of observations across hundreds of variables with dozens of latent components). In simulations, our methods recover latent components that are discarded by PCA+ICA methods. We apply our method to dependent multivariate data and demonstrate that LCA is a useful data visualization and dimension reduction tool that reveals features not apparent from PCA or PCA+ICA. We also apply our method to an experiment from the Human Connectome Project with state-of-the-art temporal and spatial resolution and identify an artifact using LCA that was missed by PCA+ICA. We present theoretical results on identifiability of the LCA model and consistency of our estimator.

Keywords: Functional Magnetic Resonance Imaging, Independent Component Analysis, Neuroimaging, Principal Component Analysis, Projection Pursuit

1 Introduction

The basic independent component analysis (ICA) model is $\mathbf{X} = \mathbf{MS}$ where \mathbf{X} is an observed vector, \mathbf{S} is a latent vector of independent random variables, and \mathbf{M} is a square matrix called the mixing matrix. It is assumed that we have a sample $\mathbf{x}_1, \dots, \mathbf{x}_V$ with corresponding latent $\mathbf{s}_1, \dots, \mathbf{s}_V$. The goal is to estimate \mathbf{M} and to recover $\mathbf{s}_1, \dots, \mathbf{s}_V$. Except for Matteson and Tsay (2013) and Stögbauer et al. (2004), ICA methodology does not directly attempt to

find components that are independent but rather components that are as non-Gaussian as possible. The principle here is that any sum of ICs will be closer to Gaussian distributed than the ICs themselves. Thus, the \mathbf{s}_v are correctly recovered if they maximize some measure of non-Gaussianity.

Transformations that maximize non-Gaussianity play a prominent role in many applications including separating audio recordings in signal processing (Bell and Sejnowski, 1995), denoising in image processing (Hyvärinen et al., 1999), face recognition in computer learning (Bartlett et al., 2002), artifact removal in electrophysiology data (Delorme et al., 2007), and estimating brain networks in cognitive neuroscience (Beckmann, 2012). We propose a novel approach for modeling non-Gaussian signals and Gaussian noise that we call Likelihood Component Analysis (LCA). Consider a sample $(\mathbf{x}_v, \mathbf{s}_v, \mathbf{n}_v), v = 1, \dots, V$, from the LCA model:

$$\mathbf{X} = \mathbf{M}_\mathbf{S}\mathbf{S} + \mathbf{M}_\mathbf{N}\mathbf{N} \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^T$; $\mathbf{S} \in \mathbb{R}^Q$ is a vector of mutually independent non-Gaussian random variables with $1 \leq Q \leq T$; $\mathbf{M}_\mathbf{S} \in \mathbb{R}^{T \times Q}$; $\mathbf{M}_\mathbf{N} \in \mathbb{R}^{T \times (T-Q)}$; $\mathbf{M} = [\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]$ (the concatenation of $\mathbf{M}_\mathbf{S}$ and $\mathbf{M}_\mathbf{N}$) is full rank; and \mathbf{N} is $(T - Q)$ -variate normal. One observes $\mathbf{x}_1, \dots, \mathbf{x}_V$ while $\mathbf{s}_1, \dots, \mathbf{s}_V$ and $\mathbf{n}_1, \dots, \mathbf{n}_V$ are latent. We assume $\mathbf{E} \mathbf{S} = \mathbf{0}$ and $\mathbf{E} \mathbf{N} = \mathbf{0}$ in (1) (in practice, data are centered by their sample mean). Our goal is to estimate $\mathbf{M}_\mathbf{S}$ and the realizations $\mathbf{s}_1, \dots, \mathbf{s}_V$ of \mathbf{S} , which we call likelihood components (LCs). By simultaneously performing dimension reduction and latent variable estimation, we will demonstrate through simulations and two real applications that estimation of the proposed model allows the discovery of non-Gaussian signals discarded by other popular methods.

Classic independent component analysis (ICA) and principal component analysis followed by ICA (hereafter, PCA+ICA) are among the most commonly used models for extracting non-Gaussian signals. Unlike (1), the classic ICA model assumes the number of components is equal to the dimension of the observations such that $\mathbf{M}_\mathbf{S}$ is square and $\mathbf{N} = \mathbf{0}$ (Hyvärinen and Oja, 2000). In practice, PCA is applied to the observations $\mathbf{x}_1, \dots, \mathbf{x}_V$ prior to classic ICA to meet the assumption of square mixing and to reduce computational costs (Hyvärinen et al., 2001). We will demonstrate that removing the smallest principal components (PCs) can discard the relevant signal (see also Green et al. 2002). Non-Gaussian signals are often discarded by PCA+ICA when they are associated with small variance. When the motivating scientific problem has a low signal-to-noise ratio, we will demonstrate that LCA is particularly well-suited to recovering the non-Gaussian signals.

PCA+ICA is commonly used to identify brain networks and artifacts in neuroimaging

(Beckmann, 2012). In fMRI, the blood oxygen level dependence (BOLD) signal is measured across time at thousands of voxels (three-dimensional analogue of a pixel) across the brain. ICA of fMRI requires dimension reduction via PCA prior to the application of ICA. It is believed that ICA can “unmix” the BOLD signal to reveal the underlying functional architecture of the brain. The existence and importance of these networks has been corroborated by other neuroimaging modalities and by the application of other statistical methods (Sporns, 2011). Additionally, ICA is commonly used for artifact removal in electroencephalography (EEG) and fMRI. Estimated independent components (ICs) that correspond to physiological noise and/or motion are identified, and accounting for these artifacts can improve subsequent analyses (Griffanti et al., 2014; Delorme et al., 2007). Even though the results from the two-stage PCA+ICA approach have been useful in the applied sciences, we show in an example that a single analysis that uses non-Gaussianity for both dimension reduction and extracting LCs can provide novel insight.

As an alternative to classic ICA, the noisy-ICA model posits that the number of noise components is equal to the dimension of the data and typically assumes isotropic noise: $\mathbf{M}_\mathbf{N}\mathbf{N} \sim N(0, \sigma^2\mathbf{I}_T)$, where \mathbf{I}_T is the $T \times T$ identity matrix. Beckmann and Smith (2004) propose a variant of PCA+ICA as an approximation to the noisy-ICA model, where they estimate the number of ICs and achieve dimension reduction using probabilistic PCA (Tipping and Bishop, 1999). Alternatively, independent factor analysis (IFA) could be used for simultaneous dimension reduction and latent variable estimation wherein the ICs are modeled as Gaussian mixtures (Attias, 1999). Allasonniere and Younes (2012) developed stochastic EM algorithms to estimate the IFA model and proposed a number of plausible parametric methods. Nonetheless, it is difficult to apply IFA to moderately sized datasets. Letting m denote the number of elements in each Gaussian mixture and Q the number of non-Gaussian components, an m^Q -dimensional integral must be approximated at each iteration, which quickly becomes computationally intractable (Allasonniere and Younes, 2012). Guo and Tang (2013) developed a multi-subject IFA model, although their application to big data utilizes PCA. Amato et al. (2010) develop non-parametric density estimators of the component densities in the noisy-ICA model but assume $\mathbf{M}_\mathbf{S}$ is semi-orthogonal (has orthogonal columns), which is not realistic for our application.

There are a number of other methods that explore structure in multivariate data using non-Gaussianity. Non-Gaussian measures of information such as kurtosis were first explored in projection pursuit algorithms (Huber, 1985), which sequentially extract “interesting” directions of information using a fixed projection pursuit index. Miettinen et al. (2014) developed the deflationary FastICA algorithm to adaptively select a parametric projection pursuit index for each non-Gaussian direction; however, their method assumes $Q = T$. Non-

Gaussian component analysis (NGCA) is a variant of independent subspace analysis (Theis, 2006) which represents data using a Gaussian subspace and an independent non-Gaussian subspace, and the non-Gaussian subspace is estimated using multiple projection pursuit indices or radial basis functions (Kawanabe et al., 2007). However, NGCA does not model independent components, and thus does not lend itself to identifying brain networks and/or artifacts. The ICA model for the case when the number of components is less than the dimension of \mathbf{X} is sometimes called under-complete ICA (Amari, 1999) or over-determined ICA, and can be estimated using a natural gradient descent algorithm. Overall, the LCA model is unique in that it specifies a generative model for the non-Gaussian signal while also defining a subspace containing Gaussian noise.

In this paper, we present a method for simultaneous dimension reduction and latent variable extraction that uncovers features that are not detected using current models. In Section 2, we define conditions for the identifiability of the LCA model in (1) and propose parametric and semi-parametric estimators. In Section 3, we investigate simulations when the observations of the latent variables are independently and identically distributed. In Section 4, we examine model robustness by applying our method to temporally and spatially structured simulated data. In Section 5, we use LCA for data visualization and dimension reduction in multivariate data. In Section 6, we estimate brain networks and artifacts from high-resolution fMRI data from the Human Connectome Project. In Section 7, we present our conclusions and discuss avenues for future research. Proofs and supplemental material are in the Appendix.

2 Methodology

2.1 Model identifiability

The identifiability of the LCA model can be established using the theorem on the uniqueness of decomposition of the “linear structure model” described in Kagan et al. (1973). Throughout this section we will assume (for simplicity) all random variables are mean zero. Define the equivalence relation $\mathbf{B} \cong \mathbf{C}$ if \mathbf{B} equals \mathbf{C} up to scaling and permutation of columns. Let “ $\stackrel{d}{=}$ ” denote equality in distribution. The following theorem can be established using Theorem 10.3.9 in Kagan et al. (1973).

Theorem 1. *Let $\mathbf{X} = \mathbf{M}_\mathbf{S}\mathbf{S} + \mathbf{M}_\mathbf{N}\mathbf{N}$, where $\mathbf{M}_\mathbf{S} \in \mathbb{R}^{T \times Q}$, the elements of \mathbf{S} are mutually independent non-Gaussian components, $\mathbf{M}_\mathbf{N} \in \mathbb{R}^{T \times (T-Q)}$, \mathbf{N} is $(T-Q)$ -variate normal, and $[\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]$ is full rank. Then for any other representation $\mathbf{X} = \mathbf{M}_\mathbf{S}^*\mathbf{S}^* + \mathbf{E}^*$ where $\mathbf{S}^* \in \mathbb{R}^Q$ are independent non-Gaussian components and \mathbf{E}^* is multivariate normal, we have: $\mathbf{M}_\mathbf{S}^* \cong \mathbf{M}_\mathbf{S}$;*

$\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations; and $\mathbf{E}^* \stackrel{d}{=} \mathbf{M}_\mathbf{N} \mathbf{N}$.

All proofs appear in the Appendix.

From Theorem 1, the signal, $\mathbf{M}_\mathbf{S} \mathbf{S}$, has a unique decomposition (on the equivalence class of scalings and permutations) into a fixed matrix and independent components. Note that the noise, $\mathbf{M}_\mathbf{N} \mathbf{N}$, does not have a unique decomposition (e.g., if \mathbf{N} comprises independent normals with equal variance, then $\mathbf{M}_\mathbf{N} \mathbf{O}'$ and $\mathbf{O} \mathbf{N}$ for orthogonal \mathbf{O} is another decomposition with independent components). Let $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_Q]'$. We state the assumptions of the LCA model below.

Assumption 1. *Assume that the model in (1) holds with*

(i) $\mathbf{S}_1, \dots, \mathbf{S}_Q$ mutually independent, non-Gaussian random variables with $\mathbf{E} \mathbf{S} = \mathbf{0}$ and $\mathbf{E} \mathbf{S} \mathbf{S}^T = \mathbf{I}_Q$.

(ii) $\text{rank}([\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]) = T$.

(iii) \mathbf{N} is $(T - Q)$ -variate (non-degenerate) normal with $\mathbf{E} \mathbf{N} = \mathbf{0}$.

Without loss of generality, we will assume that \mathbf{N} is standard multivariate normal. Let f_1, \dots, f_Q be the true densities of the LCs (the signal components). For the purposes of this paper, we will also assume f_1, \dots, f_Q are absolutely continuous, although identifiability holds more generally. Denote the eigenvalue decomposition (EVD) of the covariance matrix of \mathbf{X} by $\Sigma = \mathbf{U} \Lambda \mathbf{U}'$. Let $\mathbf{L} = \mathbf{U} \Lambda^{-1/2} \mathbf{U}'$ be a whitening matrix (the covariance matrix of $\mathbf{L}^{-1} \mathbf{X}$ is \mathbf{I}_T), and define $\mathbf{W} = \mathbf{M}^{-1} \mathbf{L}^{-1}$ where $\mathbf{M} = [\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]$. Note that $\mathbf{W} \in \mathcal{O}_{T \times T}$, where $\mathcal{O}_{T \times T}$ is the class of $T \times T$ orthogonal matrices. Let \mathbf{w}'_q denote the q th row of \mathbf{W} , and let $\mathbf{W}_\mathbf{S}$ denote the first q rows. Let $\phi(x)$ denote the standard normal density. Noting that $|\det \mathbf{W}| = 1$, we have

$$f_{\mathbf{X}}(\mathbf{x} | \mathbf{W}, \mathbf{L}) = \det(\mathbf{L}) \prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{L} \mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L} \mathbf{x}). \quad (2)$$

Note that for a density and its corresponding row of the mixing matrix, $\{f_q, \mathbf{w}_q\}$, we can trivially define a density $f_q^*(x) = f_q(-x)$ and vector $\mathbf{w}_q^* = -\mathbf{w}_q$ such that $f_q^*(\mathbf{w}_q^{*'} \mathbf{x}) = f_q(\mathbf{w}_q' \mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^T$. In this sense, we say the density and vector pair, $\{f_q, \mathbf{w}_q\}$, is identifiable up to sign. We can now establish the identifiability of the LCA model.

Corollary 1. *Suppose the linear structure model in (1) with density defined in (2) and suppose that Assumption 1 holds. Then $\{f_1, \mathbf{w}_1\}, \dots, \{f_Q, \mathbf{w}_Q\}$ are identifiable up to sign and ordering. Note the rows \mathbf{w}_{Q+k} for $k = 1, \dots, T - Q$ are not identifiable.*

2.2 The general LCA estimator

Now let $\mathbf{x}_1, \dots, \mathbf{x}_V$ be an iid sample of \mathbf{X} . Since $E\mathbf{X} = \mathbf{0}$, we will demean the data so that $\sum_{v=1}^V \mathbf{x}_v = \mathbf{0}$, and assume such hereafter. Let $\mathbb{R}_+^{T \times T}$ denote the set of $T \times T$ positive definite matrices. Assume $V \geq T$. Let $\hat{\Sigma}$ be the sample covariance matrix of $\mathbf{x}_1, \dots, \mathbf{x}_V$. Consider its eigenvalue decomposition, $\hat{\Sigma} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}'$. Then define $\hat{\mathbf{L}} = \hat{\mathbf{U}}\hat{\Lambda}^{-1/2}\hat{\mathbf{U}}'$. Note that $\sum_{v=1}^V \mathbf{o}'_q \hat{\mathbf{L}}\mathbf{x}_v = 0$. We have

$$\sum_{v=1}^V \log \phi(\mathbf{o}'_q \hat{\mathbf{L}}\mathbf{x}_v) = -\frac{V}{2} (\log 2\pi + 1). \quad (3)$$

Let $\mathcal{O}_{Q \times T}$ be the class of $Q \times T$ semi-orthogonal matrices. Define the estimator

$$\widehat{\mathbf{W}}_{\mathbf{S}}^{Tr} = \underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \sum_{v=1}^V \sum_{q=1}^Q \log f_q(\mathbf{o}'_q \hat{\mathbf{L}}\mathbf{x}_v). \quad (4)$$

The superscript “Tr” indicates that $\widehat{\mathbf{W}}_{\mathbf{S}}^{Tr}$ is computed when Q and the true component densities are known, so $\widehat{\mathbf{W}}_{\mathbf{S}}^{Tr}$ is an oracle estimator that cannot be used in practice. In our Logis-LCA estimator, f_1, \dots, f_Q are replaced by the logistic density, while in Spline-LCA we alternate between approximating f_1, \dots, f_Q by tilted Gaussians with $\widehat{\mathbf{W}}_{\mathbf{S}}$ fixed and computing $\widehat{\mathbf{W}}_{\mathbf{S}}$ by (4) with f_1, \dots, f_Q fixed at estimates.

Observe that the problem of estimating $\mathbf{W}_{\mathbf{S}}$ is equivalent to the problem of estimating the LCs because $\hat{\mathbf{s}}_v = \widehat{\mathbf{W}}_{\mathbf{S}} \hat{\mathbf{L}}\mathbf{x}_v$ for all v . Thus we would like a consistent estimator of $\mathbf{W}_{\mathbf{S}}$. Towards this, we have the following proposition.

Proposition 1. *Consider a random vector $\mathbf{Y} \in \mathbb{R}^T$ with density $f_{\mathbf{Y}}$ such that $E\mathbf{Y} = \mathbf{0}$ and $E\mathbf{Y}\mathbf{Y}' = \mathbf{I}_T$. Then for any \mathbf{o} and \mathbf{w} such that $\mathbf{o}'\mathbf{o} = \mathbf{w}'\mathbf{w} = 1$, we have*

$$E \log \phi(\mathbf{o}'\mathbf{Y}) = E \log \phi(\mathbf{w}'\mathbf{Y}).$$

Next, we show the consistency of the oracle estimate of $\mathbf{W}_{\mathbf{S}}$.

Theorem 2. *Suppose \mathbf{X} follows the LCA model in (1) with Assumption 1 holding and assume the non-Gaussian components have bounded absolutely continuous densities (satisfied by the classes considered below). Additionally assume $E\mathbf{X} = \mathbf{0}$ and $E\mathbf{X}\mathbf{X}' = \mathbf{I}$. Let $\mathbf{W}_{\mathbf{S}}$ denote the first Q rows of \mathbf{M}^{-1} . Given an iid sample $\mathbf{x}_1, \dots, \mathbf{x}_V$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Tr} \rightarrow \mathbf{W}_{\mathbf{S}}$ almost surely on the equivalence class of signed permutations.*

Since $\widehat{\mathbf{W}}_{\mathbf{S}}$ is not invertible, we also define an estimator of $\mathbf{M}_{\mathbf{S}}$:

$$\widehat{\mathbf{M}}_{\mathbf{S}} = \underset{\mathbf{M} \in \mathbb{R}^{T \times Q}}{\operatorname{argmin}} \sum_{v=1}^V \|\mathbf{x}_v - \mathbf{M} \hat{\mathbf{s}}_v\|_2^2. \quad (5)$$

Although we assume iid observations in the construction of (4), the LCA model is capable of recovering many forms of dependent data, as is also the case in ICA. This will be demonstrated in simulations.

There is a natural ordering of the LCs when the component densities are not equal. Additionally, if we assume finite third moments and the component densities are skewed, we can assume positive skewness and then the LCA model is fully identifiable (as in ICA, Eloyan and Ghosh 2013). We define a canonical ordering for a sample $\mathbf{x}_1, \dots, \mathbf{x}_V$:

$$\sum_{v=1}^V f_1(\mathbf{w}'_1 \mathbf{L} \mathbf{x}_v) > \sum_{v=1}^V f_2(\mathbf{w}'_2 \mathbf{L} \mathbf{x}_v) > \dots > \sum_{v=1}^V f_Q(\mathbf{w}'_Q \mathbf{L} \mathbf{x}_v).$$

with

$$\sum_{v=1}^V (\mathbf{w}'_q \mathbf{L} \mathbf{x}_v)^3 > 0$$

for $q = 1, \dots, Q$. In practice, we force the sample third moment to be positive and order components by their likelihoods.

2.3 A sign- and permutation-invariant discrepancy measure for non-square matrices

To assess the accuracy of our estimates and/or compare multiple estimates, we need a discrepancy measure that is invariant on the equivalence class of signed permutation matrices, and we would like a measure that can apply to matrices of differing dimensions when the estimated number of components may not equal Q . We cannot use the Amari or the minimum distance (Ilmonen et al., 2010) measures because $\mathbf{M}_{\mathbf{S}}$ is non-square. We propose a novel measure of dissimilarity that uses the Hungarian algorithm to match rows of the unmixing matrix as in Risk et al. (2014) but applies to non-square unmixing. We also generalize the measure to apply to matrices that may have a different number of columns, in which case the measure only compares matching columns. In addition to diagnosing accuracy, this measure will also be used to assess convergence in our algorithms.

Consider $\mathbf{M}_1 \in \mathbb{R}^{T \times Q}$ and $\mathbf{M}_2 \in \mathbb{R}^{T \times R}$ with $Q \leq R$. Let \mathcal{P}_{\pm} be the class of $R \times Q$ signed

permutation matrices, so that post-multiplication of \mathbf{M}_2 by $\mathbf{P}_\pm \in \mathcal{P}_\pm$ results in a subset of Q (permuted) columns of \mathbf{M}_2 for $Q < R$. Define the sign- and permutation-invariant mean-squared error:

$$PMSE(\mathbf{M}_1, \mathbf{M}_2) = \operatorname{argmin}_{\mathbf{P}_\pm \in \mathcal{P}_\pm} \|\mathbf{M}_1 - \mathbf{M}_2 \mathbf{P}_\pm\|_F^2, \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm and the optimal \mathbf{P}_\pm is found using the Hungarian algorithm. In practice, we also standardize the columns of \mathbf{M}_1 and \mathbf{M}_2 to have unit norm, and thus the measure is scale invariant. Another advantage of this measure is that it can be used to compare independent components directly. If \mathbf{S}_1 is a $V \times Q$ matrix in which each row is a realization of the LC in \mathbb{R}^Q , and if $\mathbf{S}_2 \in \mathbb{R}^{V \times R}$, then we define their discrepancy as $PMSE(\mathbf{S}_1, \mathbf{S}_2)$.

2.4 A parametric model: Logis-LCA

First we present a parametric method called Logis-LCA in which the densities of the LCs are approximated by logistic densities. The logistic density is used in the Infomax algorithm, where it appears to work well for unmixing audio signals (Bell and Sejnowski, 1995) and brain networks (Correa et al., 2007). Under the constraint of zero mean and unit variance, the logistic density has the form

$$f(x) = \frac{\exp(-x/\frac{\sqrt{3}}{\pi})}{\frac{\sqrt{3}}{\pi} \left\{ 1 + \exp(-x/\frac{\sqrt{3}}{\pi}) \right\}^2}. \quad (7)$$

We define our estimator for some $Q^* \leq T$ such that Q^* may or may not equal the true number of LCs, Q . Applying (7) to (4) and the centered data $\mathbf{x}_1, \dots, \mathbf{x}_V$, the Logis-LCA estimator of $\mathbf{W}_\mathbf{S}$ can be written as

$$\widehat{\mathbf{W}}_\mathbf{S}^L = \operatorname{argmax}_{\mathbf{O}_\mathbf{S} \in \mathcal{O}_{Q \times T}} - \sum_{v=1}^V \sum_{q=1}^{Q^*} \log \left\{ 1 + \exp \left(-\mathbf{o}'_q \widehat{\mathbf{L}} \mathbf{x}_v \frac{\pi}{\sqrt{3}} \right) \right\}. \quad (8)$$

We maximize (8) using a modification of the symmetric fixed-point ICA algorithm (Hyvarinen, 1999). ICA implementations require the estimator to be a square matrix. We orthogonalize intermediate estimates of the rows of $\mathbf{W}_\mathbf{S}^L$ by calculating the SVD and setting the singular values equal to one. Additionally, we assess convergence using (6). See Section C in the Appendix for details.

2.5 A semi-parametric model: Spline-LCA

In this section, we use the flexible family of tilted Gaussian densities to model the LCs. The proposed model is equivalent to ProDenICA (Hastie and Tibshirani, 2003) when $Q = T$. For $Q < T$, it can be shown that the likelihood extends the semiparametric likelihood in Blanchard et al. (2006) to include an independence model for the LCs (see Appendix). The independence assumption is necessary for physically and biologically useful interpretations.

Suppose the LCs have tilted Gaussian distributions of the form $\phi(x)e^{g(x)}$, where the tilt function, $g(x)$, is a twice-differentiable function. Define the log-likelihood for some $\mathbf{O} \in \mathcal{O}_{T \times T}$:

$$\begin{aligned} \ell(\mathbf{O}, g_1, \dots, g_{Q^*} \mid \hat{\mathbf{L}}, Q^*, \mathbf{x}_1, \dots, \mathbf{x}_V) \\ = \sum_{v=1}^V \left[\sum_{q=1}^{Q^*} \left\{ \log \phi(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) + g_q(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) \right\} + \sum_{k=1}^{T-Q^*} \log \phi(\mathbf{o}'_{k+Q^*} \hat{\mathbf{L}} \mathbf{x}_v) \right]. \end{aligned}$$

This log-likelihood does not have an upper bound, so we define a penalized log-likelihood:

$$\begin{aligned} \ell_{pen}(\mathbf{O}, g_1, \dots, g_{Q^*} \mid \hat{\mathbf{L}}, Q^*, \mathbf{x}_1, \dots, \mathbf{x}_V) = & - \sum_{q=1}^{Q^*} \lambda_q \int \{g''_q(x)\}^2 dx - \int \phi(x) e^{g_q(x)} dx \quad (9) \\ & + \frac{1}{V} \sum_{v=1}^V \sum_{q=1}^{Q^*} g_q(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) - \frac{T}{2} (\log 2\pi + 1), \end{aligned}$$

where we have used (3) to simplify the Gaussian components.

Now consider the problem of estimating $g(x)$ for fixed \mathbf{O} .

Proposition 2. *Let G be the class of all cubic splines $g : \mathbb{R} \rightarrow \mathbb{R}$. Consider the argmax of (9) for $g_q \in G$. Then (i) $\int \phi(x) e^{g_q(x)} dx = 1$ and (ii) $\int x \phi(x) e^{g_q(x)} dx = 0$ for each q .*

We adapt the ProDenICA algorithm of Hastie and Tibshirani (2003) to LCA, in which we alternate between estimating \mathbf{W}_S for fixed $\hat{f}_1, \dots, \hat{f}_{Q^*}$ via the fixed point algorithm and estimating f_1, \dots, f_{Q^*} for fixed $\hat{\mathbf{W}}_S$ using the ‘‘Poisson trick’’. Our account largely follows the description in Hastie et al. (2009) but for semi-orthogonal (rather than orthogonal) matrices.

Suppose \mathbf{W}_S is given and define $s_{vq} = \mathbf{w}'_q \mathbf{z}_v$, where $\mathbf{z}_v = \hat{\mathbf{L}} \mathbf{x}_v$. Let x_1^*, \dots, x_{L+1}^* define a discretization, $[x_1^*, x_2^*), [x_2^*, x_3^*), \dots, [x_L^*, x_{L+1}^*)$, of the support of the tilt function of the non-Gaussian densities such that $\Delta = x_\ell^* - x_{\ell-1}^*$ for all $\ell = 2, \dots, L+1$. It suffices to take $x_1^* = \min(s_{11}, \dots, s_{nd}) - 0.1\hat{\sigma}_z$ and $x_{L+1}^* = \max(s_{11}, \dots, s_{nd}) + 0.1\hat{\sigma}_z$, where $\hat{\sigma}_z$ denotes the sample standard deviation, which here is equal to one. Next, let $x_\ell = \frac{1}{2}(x_\ell^* + x_{\ell+1}^*)$. For each $q \in \{1, \dots, Q^*\}$ and $\ell \in \{1, \dots, L\}$, define

$$y_{\ell q} = \sum_{v=1}^V \mathbb{1}\{s_{vq} \in [x_{\ell}^*, x_{\ell+1}^*)\}.$$

We approximate (9) by discretizing the first integral and estimating the sum over V as a weighted sum over L . Restricting our attention to a single q , we have

$$-\lambda_q \int \{g_q''(x)\}^2 dx + \sum_{\ell=1}^L \left[\frac{y_{\ell q}}{V} \{g_q(x_{\ell}) + \log \phi(x_{\ell})\} - \Delta \phi(x_{\ell}) e^{g_q(x_{\ell})} \right].$$

Dividing by Δ , we have

$$-\beta_q \int \{g_q''(x)\}^2 dx + \sum_{\ell=1}^L \left[\frac{y_{\ell q}}{V\Delta} \{g_q(x_{\ell}) + \log \phi(x_{\ell})\} - \phi(x_{\ell}) e^{g_q(x_{\ell})} \right] \quad (10)$$

for some penalty β_q . This is proportional to a Poisson generalized additive model (GAM), where $\frac{y_{\ell q}}{V\Delta}$ is the response and the expected response is equal to $\phi(x_{\ell}) e^{g_q(x_{\ell})}$. This can be fit using the **gam** package in R (Hastie, 2013) where β_q is chosen to result in a user-specified number of effective degrees of freedom. We find that $df = 8$ and $L = 100$ produce fast and accurate density estimates in simulations for a variety of densities when the sample size V is equal to 1,000. This method also easily scales to tens of thousands of observations.

The algorithm to estimate both $\mathbf{W}_{\mathbf{S}}$ and f_1, \dots, f_Q is summarized below. Note that step 3 requires the first and second derivatives of the log densities of the LCs, which makes the use of B-splines convenient.

Algorithm 1: The Spline-LCA algorithm.

Inputs : The whitened $V \times T$ data matrix \mathbf{Z} ; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ .

Result: Estimates of the latent components, $\hat{\mathbf{S}}$, and their densities, $\hat{f}_1, \dots, \hat{f}_Q$.

1. Let $n = 0$ and define $\mathbf{S}^{(n)} = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{(n)'}.$
 2. Estimate $f_q^{(n+1)}$ for $q = 1, \dots, Q$.
 3. Using (4), update $\mathbf{W}_{\mathbf{S}}^{(n+1)}$ given $f_1^{(n+1)}, \dots, f_Q^{(n+1)}$ and $\mathbf{S}^{(n)}$ with one-step of the fixed-point algorithm (see Appendix).
 4. Let $\mathbf{S}^{(n+1)} = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{(n+1)'}$.
 5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(n+1)'}, \mathbf{W}_{\mathbf{S}}^{(n)'}) < \epsilon$, stop, else increment n and repeat (2)-(4).
-

3 Simulations: Distributional and Noise-rank Assumptions

In this section, we simulate the LCA model [given by (1) with $\mathbf{M}_S \in \mathbb{R}^{T \times Q}$] and the noisy-ICA model [again given by (1) with $\mathbf{M}_S \in \mathbb{R}^{T \times Q}$ but now with $\mathbf{M}_N \mathbf{N} \sim N(0, \sigma^2 \mathbf{I}_T)$] under a variety of source distributions in which the components are iid as well as a scenario in which the sources are sparse images. We compare (i) deflationary FastICA with the ‘log cosh’ nonlinearity (D-FastICA), where the deflation option estimates components one-by-one such that the algorithm is considered a projection pursuit method (Hyvärinen and Oja, 2000); (ii) two-class IFA with isotropic noise (IFA); (iii) PCA followed by Infomax (PCA+Infomax); (iv) PCA followed by ProDenICA (PCA+ProDenICA) (v) Logis-LCA; and (vi) Spline-LCA. We evaluate the robustness of these methods with respect to assumptions on the rank of the noise components, distribution of the latent components, and the signal-to-noise ratio (SNR). We define the SNR as the ratio of the total variance from the mixed non-Gaussian components to the total variance from the noise components. Formally, consider the non-zero eigenvalues $\lambda_1, \dots, \lambda_Q$ from the covariance matrix of $\mathbf{M}_S \mathbf{S}$. For LCA, let $\lambda_{\epsilon_1}, \dots, \lambda_{\epsilon_{T-Q}}$ denote the eigenvalues from the EVD of the covariance matrix of $\mathbf{M}_N \mathbf{N}$. Then,

$$SNR = \frac{\sum_{q=1}^Q \lambda_q}{\sum_{k=1}^{T-Q} \lambda_{\epsilon_k}}. \quad (11)$$

For the noisy-ICA model, we have T non-zero eigenvalues in the denominator sum.

We fit D-FastICA using the fastICA R package (Marchini et al., 2010). We fit PCA+Infomax using our own implementation of the Infomax algorithm. We fit PCA+ProDenICA using the ProDenICA function from the R package of that name (Hastie and Tibshirani, 2010). Note that these methods can provide an estimate of \mathbf{S} but not the mixing matrix, which we estimated using (5). We fit the IFA model with two-component mixtures of normals using our own implementation, and the ICs were estimated by their conditional means (see equation (81) in Attias 1999). Details are in the Appendix.

3.1 Simulation Design

Data were generated with $T = 5$ and $Q = 2$ according to a $2^2 \times 6$ full factorial design. The three factors were

- i) **The model:** the levels were (a) the LCA model with rank- $(T - Q)$ noise and (b) the noisy-ICA model with rank- T noise. In both models the signal was $\mathbf{M}_S \mathbf{S}$ where \mathbf{M}_S is $T \times Q$ with $Q < T$.

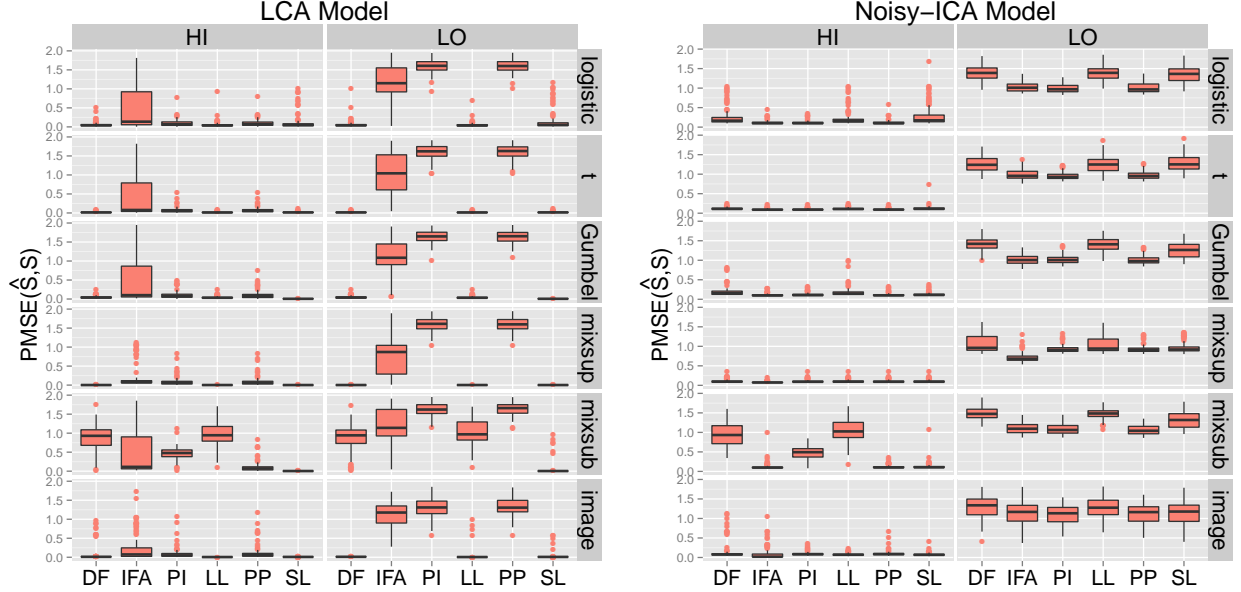
- ii) **The signal to noise (SNR) ratio:** the levels were (a) high where the ratio of the variance from the signal components to the variance from the noise components was 5:1 and (b) low where that ratio was 1:5.
- iii) **Signal distribution:** the levels were (a) logistic, (b) t, (c) Gumbel, (d) sub-Gaussian mixture of normals, (e) super-Gaussian mixture of normals, (f) with values determined by a sparse image, as described below. The two signal components were each iid and had the same distributions in cases (a)–(e) but different distributions in the sparse signal case.

Since we generated $Q = 2$ signal components for all simulations, there were $T - Q = 3$ and $T = 5$ noise components for the LCA model and noisy-ICA model, respectively. Observations in the noise components were iid isotropic normal except for the sparse image scenario, in which we used the R-package neuRosim (Welvaert et al., 2011) to generate three-dimensional Gaussian random fields with full width at half maximum (FWHM) equal to 6 for each noise component.

The signal components had scale parameter equal to $\sqrt{3}/\pi$ for the logistic, 5 degrees of freedom for the t, and scale parameter equal to $\sqrt{6}/\pi$ for the Gumbel. For the super-Gaussian mixture of normals, we simulated a two-class model with the first centered at 0 with variance 2/3 with probability 0.95 and the second centered at 5 with unit variance (excess kurtosis ≈ 9), which is motivated by a brain network with 5% of voxels activated. For the sub-Gaussian mixture of normals, we used the two-class model with the first centered at -1.7 with unit variance and probability 0.75 and the second centered at 1.7 with unit variance and probability equal to 0.25, which is equivalent to distribution ‘l’ from Hastie and Tibshirani (2003) (excess kurtosis ≈ -0.3). For the sparse image, we used neuRosim to generate a three-dimensional image in which all voxels were iid normal with variance equal to 0.0001 except, in the first component, a sphere of radius two in which the center was located at $(5, 5, 5)$ with voxel-value equal to one and the exponential decay rate set to 0.5. The second sparse image component was similar except the feature was a cube centered at $(7, 7, 7)$ with width equal to two and exponential decay rate equal to one.

We conducted 112 simulations (chosen because we used a cluster with 56 processors) with $V = 1,000$ observations in which \mathbf{M}_S and \mathbf{M}_N were randomly generated to have condition number between one and ten for each combination of factors. Since neither the set of orthogonal matrices (PCA+ICA methods) nor semi-orthogonal matrices (LCA methods) is convex, we approximated the argmax by initializing D-FastICA, PCA+Infomax, Logis-LCA, PCA+ProDenICA, and Spline-LCA from twenty random matrices and selecting the estimate associated with the largest objective function value. For Logis-LCA and Spline-LCA, ten of

Figure 1: Boxplots of $PMSE$ for estimated columns of \mathbf{S} where the rank of the noise was $T-Q$ (LCA Model) or T (Noisy-ICA Model) in high SNR ('HI') and low SNR ('LO') scenarios for various latent distributions. 'DF' = D-FastICA; 'IFA' = independent factor analysis; 'PI' = PCA+Infomax; 'LL' = Logis-LCA; 'PP' = PCA+ProDenICA; 'SL' = Spline-LCA.



these twenty initializations were from random matrices constrained to the principal subspace. Let $\hat{\mathbf{U}}_{1:Q}$ denote the first Q rows from $\hat{\mathbf{U}}$ in the decomposition $\hat{\Sigma} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}'$. Then constraining the initial matrix, \mathbf{W}_S^0 , to the principal subspace is equivalent to $\mathbf{W}_S^0 = \hat{\mathbf{U}}_{1:Q}\mathbf{O}$ with $\mathbf{O} \in \mathbf{O}_{Q \times Q}$. For IFA, one must specify initial values for the unmixing matrix, the variance of the isotropic noise, and the parameters of the Gaussian mixtures, and here we had four strategies to find the argmax. See Section D of the Appendix for additional details.

3.2 Results

When the LCA model was true and there was a high SNR, all methods generally produced accurate estimates of \mathbf{S} for the logistic, t , Gumbel, super-Gaussian mixture of normals, and sparse images, but only Spline-LCA was accurate for the sub-Gaussian mixture of normals, and the performance of IFA was more variable than other methods for all distributions (Figure 1). In these simulations, boxplots examining the accuracy of $\hat{\mathbf{M}}_S$ showed patterns similar to those found in Figure 1 and consequently are not presented.

When the LCA model was true and there was a low SNR, IFA, PCA+Infomax, and PCA+ProDenICA failed to recover the LCs for all distributions, while D-FastICA and Logis-LCA recovered all distributions except for the sub-Gaussian mixture of normals. Spline-LCA

was the method most robust to distributional assumptions and was the only method that recovered the sub-Gaussian mixture.

However, when the noisy-ICA model was true and there was a high SNR, all methods generally produced accurate estimates for the logistic, t, Gumbel, super-Gaussian, and sparse image. IFA and Spline-LCA were the only methods that recovered ICs with sub-Gaussian distributions. When the noisy-ICA model was true and there was a low SNR, all methods performed poorly, although IFA, PCA+Infomax, and PCA+ProDenICA outperformed LCA algorithms for some distributions.

Overall, LCA methods were robust to the SNR for rank- $(T - Q)$ noise, and performed well in the high SNR scenario for rank- T noise. Additionally, Spline-LCA was most robust to distributional assumptions. In contrast, IFA, PCA+Infomax, and PCA+ProDenICA performed poorly in the low SNR scenario for both the rank- $(T - Q)$ and rank- T noise.

4 Simulations: Spatio-temporal Networks

Next, we examine the ability of D-FastICA, PCA+Infomax, Logis-LCA, and Spline-LCA to recover simulated networks whose loadings vary deterministically with time in the presence of spatially and temporally correlated noise, so that the simulations resemble the structure found in task-based fMRI. We also examine the effect of using $Q^* \neq Q$ on network recovery. In this way, we assess whether the LCA algorithm can recover brain networks and their temporal loadings from spatiotemporal neuroimaging. We did not include IFA in these simulations because it was difficult to estimate the mixing matrix when T was relatively large (e.g., $T = 50$). Additionally, IFA, PCA+Infomax, and PCA+ProDenICA produced similar results for most distributions in the previous simulations, and PCA+Infomax and PCA+ProDenICA were more accurate than IFA in the high SNR scenario for the LCA model. Hence, our previous simulations suggest there would be little insight gained from including IFA.

4.1 Simulation Design

We simulated three networks mixed across fifty time units. The networks were 33×33 images where “active” pixels were in the shape of a “1”, “2 2”, or “3 3 3” with values between 0.5 and 1 and “inactive” pixels were mean zero iid normal with variance equal to 0.0001 (see Figure 2). Let \mathbf{m}_q denote the q th column of \mathbf{M}_S . To simulate the temporal activation patterns of brain networks, we used neuRosim to convolve the canonical hemodynamic response function (HRF) with a block-design with a pair of onsets at $\{1, 20.6\}$, $\{10.8, 40.2\}$, and $\{10.8, 30.4\}$

for \mathbf{m}_1 , \mathbf{m}_2 , and \mathbf{m}_3 , respectively, and duration equal to 5 time units (Welvaert et al., 2011).

In the LCA scenario, noise components were generated as forty-seven independent 33×33 Gaussian random fields with FWHM=6. Temporal correlation was introduced via the mixing matrix, in which each column of \mathbf{M}_N corresponded to an AR(1) process simulated for fifty time units with AR coefficient equal to 0.47 and unit variance, where the AR coefficient was chosen based on a preliminary analysis of the fMRI data analyzed in Section 6. Additionally, noise components were scaled such that the SNR was 0.4, which approximately equals the SNR estimated in Section 6. In the noisy ICA scenario, a 33×33 Gaussian random field with FWHM=6 was simulated for $t = 1$. Then noise components were defined recursively for $t = 2, \dots, 50$ to be equal to 0.47 times the noise at time $t - 1$ plus a realization from an independent Gaussian random field with FWHM=6.

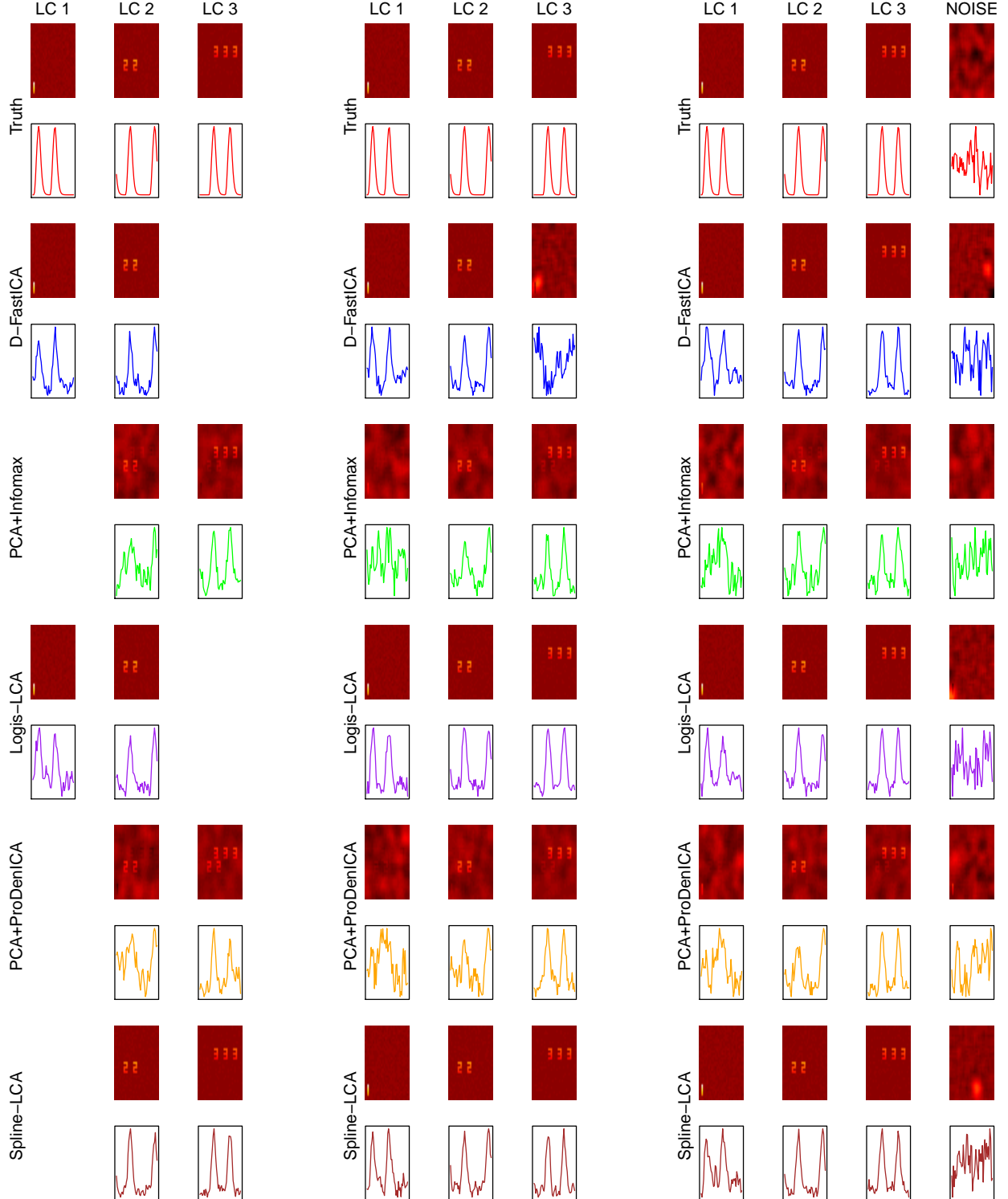
We conducted 111 simulations with $Q^* = 2, 3$ or 4 (with fixed $Q = 3$) and initialized all algorithms from twenty random mixing matrices for each simulation and each Q^* . For Logis-LCA and Spline-LCA, ten of the twenty initializations were from random matrices in the principal subspace, as in Section 3.1.

4.2 Results

By inspecting the images and loadings associated with the median $PMSE(\hat{\mathbf{S}}, \mathbf{S})$ for each method in the LCA scenario, we see that D-FastICA recovers a spurious component when $Q^* = 3$; PCA+Infomax and PCA+ProDenICA generally fail to unmix features; and Logis-LCA and Spline-LCA are highly accurate (Figure 2). It is notable that estimates from PCA+Infomax and PCA+ProDenICA were sensitive to the choice of Q^* , as when $Q^* < Q$, an estimated latent component resembled a union of components two and three. In PCA+ProDenICA, the loadings for the estimated component were highly correlated with component three ($r = 0.75$), which mistakenly suggests components two and three are functionally connected. For $Q^* = 3$, the features in the estimated component one are faintly visible in PCA+Infomax whereas component one was not recovered by PCA+ProDenICA. In contrast, Logis-LCA and Spline-LCA clearly separated components for all Q^* , such that when $Q^* < 3$, the recovered components were accurate estimates of a subset of the true ($Q = 3$) components.

For the noisy-ICA scenario, the features recovered by Logis-LCA most closely resembled the truth (Figure 3). Features from component two were again faintly visible in component three for $Q^* = 2$ in both PCA+Infomax and PCA+ProDenICA, again indicating inadequate unmixing of the networks. As seen in the LCA scenario, D-FastICA recovered a spurious component for $Q^* = 3$, but accurately estimated component three for $Q^* = 4$. Spline-LCA

Figure 2: Network recovery from the LCA scenario with $Q = 3$ for $Q^* = 2$ (first three columns), $Q^* = 3$ (columns 4-6), or $Q^* = 4$ (columns 7-10). Images depict LCs and time-series depict the loadings ($\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_{Q^*}$) corresponding to the median $PMSE(\hat{\mathbf{S}}, \mathbf{S})$ from 111 simulations. In the last column, “Truth” corresponds to an arbitrary noise component whereas the algorithms attempted to estimate a fourth LC.



was sensitive to the assumption on the rank of the noise, as it failed to recover component one, although it was quite accurate for components two and three. Spatial correlations in the noise appear to result in spurious disk-like features, which were detected in Spline-LCA and D-FastICA. For the simulation associated with the median error, an accurate estimate of component one was associated with a local maxima in Spline-LCA, but the spurious component had a higher likelihood. The component was recovered in some simulations, and the lower quartile of the PMSE error was accurate (Figure A1).

5 Data Visualization and Dimension Reduction

We used Logis-LCA and Spline-LCA for data visualization and dimension reduction in multivariate data comprising measurements from independent leaf samples (Silva et al., 2013). Fourteen variables were generated from eight to sixteen images of leaves from each of thirty species (Figure A2). Many of the covariates are highly correlated (Figure A3). We plotted the first two PCs, ICs from PCA+Infomax and PCA+ProDenICA, and LCs from Logis-LCA and Spline-LCA. Two-dimensional PCA does not reveal clear features (Figure 4). Since we are examining two dimensions, the effect of ICA is apparent as a rotation of the X- and Y-axes. Rotating the axes does not reveal any additional insight (Figure 4, Figure A4). In contrast, Spline-LCA clearly reveals three categories, where the green dots correspond to two plant species that have very thin leaves (species 31 and 34 in Figure A2), the blue category corresponds to a species with leaves that are thinner than most species but less than those comprising the green dots (species 8), and the red category corresponds to all other species. Logis-LCA also reveals structure (Figure A4), although the separation is less than in Spline-LCA.

PCA+ICA methods were sensitive to the number of components estimated whereas components were robustly estimated in the LCA methods. In PCA+Infomax and PCA+ProDenICA, the first two (matched) ICs for $Q^* = 5$ differed from the ICs estimated using two components, demonstrating the sensitivity of PCA+ICA methods to the number of principal components (Figures A4 and A5). In contrast, the first two LCs extracted from Logis-LCA and Spline-LCA when five components were estimated were very similar to the LCs estimated using two components.

6 Application to fMRI

We applied Spline-LCA to a single subject from the Social Cognition / Theory of Mind (ToM) experiment of the WU-Minn Human Connectome Project (HCP; additional information in

Figure 3: Network recovery from the noisy-ICA scenario with $Q = 3$ for $Q^* = 2$ (first three columns), 3 (columns 4-6), or 4 (columns 7-10).

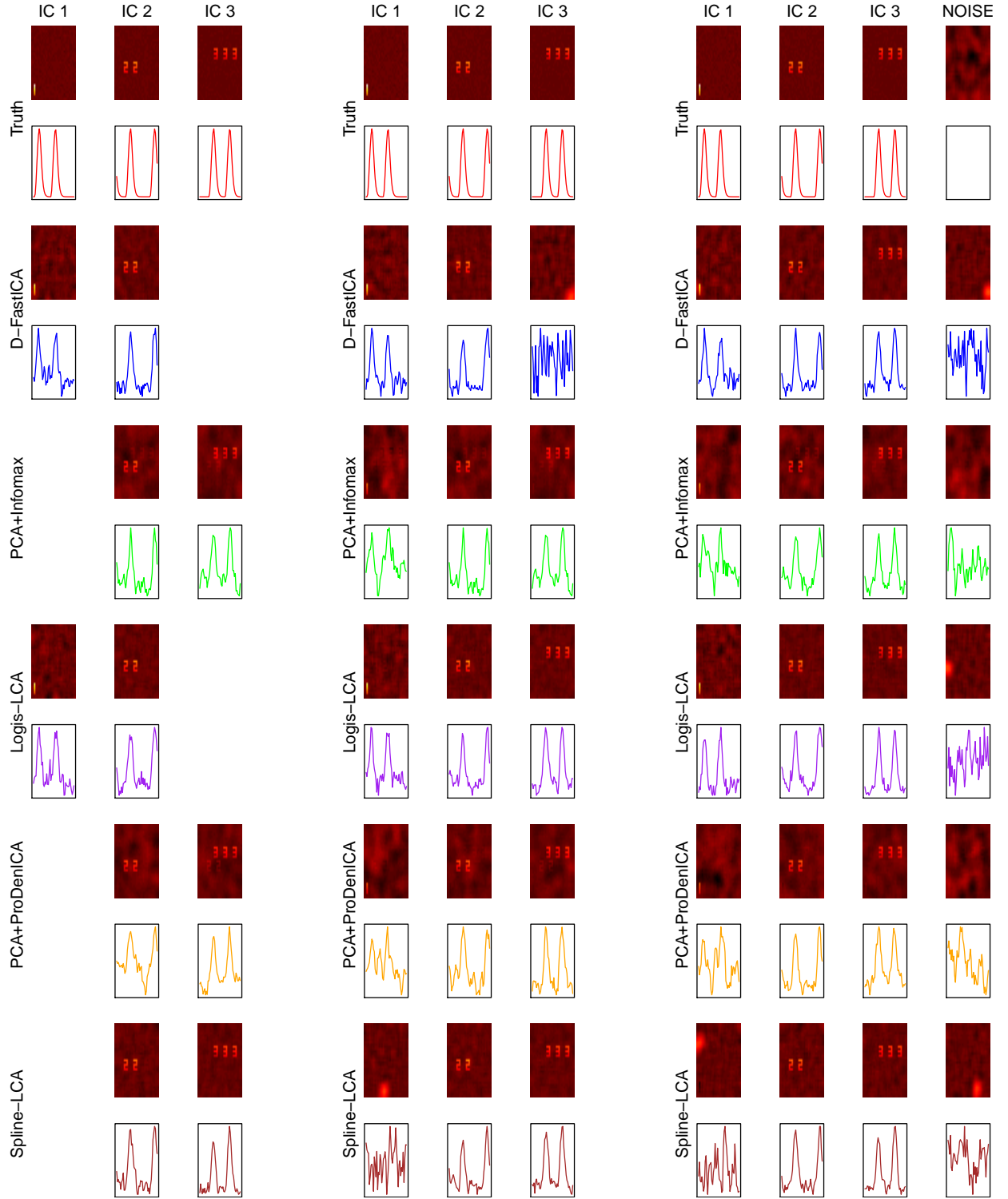
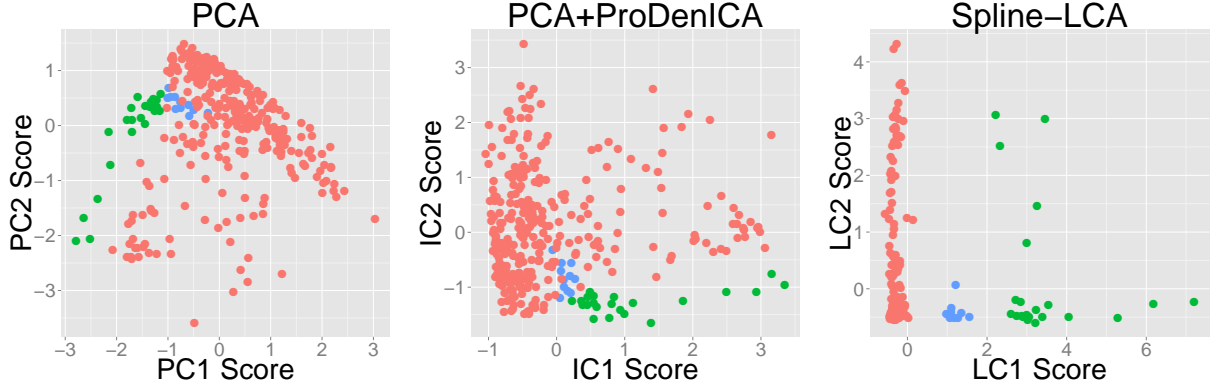


Figure 4: Data visualization and dimension reduction for the leaf dataset. The original dataset comprises 14 variables, many of which are highly correlated. The green dots correspond to *Podocarpus sp.* and *Pseudosasa japonica*; the blue dots correspond to *Neurium oleander*; the red dots correspond to all other species.



Appendix). For details of the experimental paradigm see Barch et al. (2013). We used the minimally pre-processed data (Glasser et al., 2013) from the first session of subject 103414 from the June 5, 2014, data release. Three-dimensional volume data were vectorized and non-brain tissue excluded using the mask provided from the HCP. This resulted in a $230,459 \times 272$ data matrix. Each voxel was treated as a replicate with $v = 1, \dots, V$ for $V = 230,459$, which is analogous to ‘spatial’ ICA of fMRI (Calhoun et al., 2009). We mean centered and variance normalized each voxel’s time course prior to conducting LCA, as suggested for ICA of fMRI (Beckmann and Smith, 2004).

The application of ICA to fMRI usually assumes that voxels are iid (an exception is the approach in Lee et al. 2011, which models time-series data using the Whittle likelihood). This assumption is often not made explicitly because ICA is usually derived from the perspective of maximizing non-Gaussianity. Since the objective function maximizing non-Gaussianity can also be derived from ML theory where the non-linear function is equivalent to the log likelihood (e.g., Hyvärinen and Oja 2000), summation of the non-linear function over voxels (e.g., Equation 12 in Beckmann and Smith 2004) is mathematically equivalent to assuming the voxels are independent. Despite the violation of model assumptions, ICA recovers simulated brain networks and their loadings (Beckmann and Smith, 2004) and has proven useful in constructing models of functional connectivity that are consistent across subjects and image acquisition centers (Biswal et al., 2010).

We used the ICA software MELODIC (FSL) to determine the number of components that would be used in an analogous ICA of this dataset, which estimated thirty components. We initiated the algorithm from fifty-six randomly generated matrices, twenty-eight of which were in the principal subspace. Depending on initialization, the algorithm took between ten

minutes and 3.75 hours on a 2666 MHz processor, where 3.75 hours represented initializations that reached the maximum number of iterations, which we conservatively chose to be equal to 300. We also completed an analogous PCA+ProDenICA with thirty components and fifty-six initializations using the R package ProDenICA (Hastie and Tibshirani, 2010).

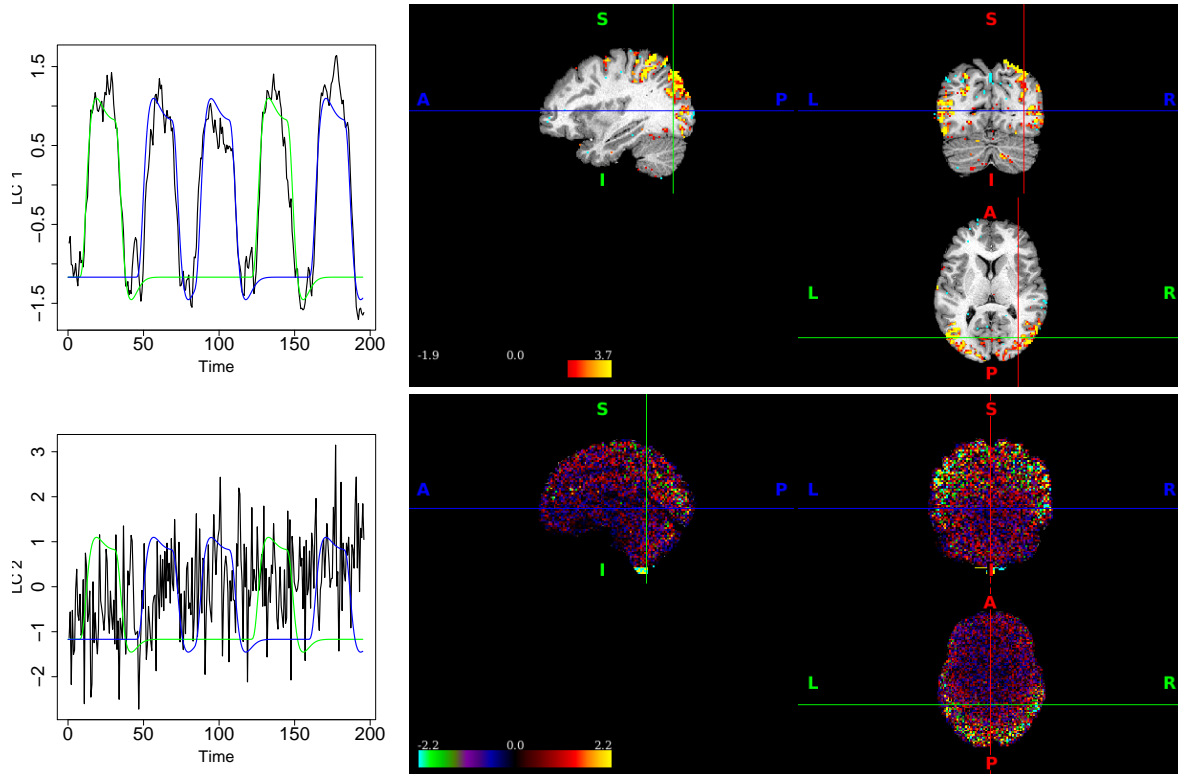
We examined the correlation between the loadings (columns of $\widehat{\mathbf{M}}_{\mathbf{S}}$) and the “mentalizing” and “random” tasks. The mentalizing and random task covariates were generated by convolving each task’s onsets and durations with the canonical HRF in SPM8 (Ashburner et al., 2004). The first component was highly correlated with the mentalize and random tasks (Figure 5). This component showed activation primarily in the lateral occipital cortex. A similar component was found using PCA+ProDenICA (not depicted).

We also detected components that were estimated in Spline-LCA but not in PCA+ProDenICA. Eight out of thirty LCs had a correlation less than 0.5 with their matched IC components. In particular, component two in Spline-LCA was not correlated with any of the components in PCA+ProDenICA (max correlation among all ICs = 0.01). This component appears to correspond to an artifact due to motion and possibly other sources of noise. Its time course was correlated with three of the motion parameters from the rigid-body alignment ($r = 0.32$, 0.32 , and 0.42 for the x-transformation, x-rotation, and z-rotation parameters, respectively). Voxels were highly activated in the brainstem, which could be due to movement. Additionally, there was a positive correlation with time ($r = 0.44$), which could be related to scanner drift. Removing artifacts from fMRI detected using ICA is a popular tool that can increase detection in subsequent mixed-modeling of voxel activation (Tohka et al., 2008). Thus, our detection of a novel artifact represents a potential benefit of LCA over current methodology.

7 Discussion

In this study, we propose a model-based method for estimating non-Gaussian latent components in the presence of Gaussian noise that has many applications including dimension reduction, signal processing, artifact detection, and network estimation. We presented two applications: data visualization and dimension reduction, and identifying brain networks and artifacts from neuroimaging. Our first simulation study indicates that our methods perform well when the LCA model is true, even for low SNR. When the noisy-ICA model is true, our methods perform well in the high SNR scenario, while none of the methods perform well in the low SNR scenario. In the second simulation study, we examined performance when data contained spatiotemporal dependence and a moderately low SNR. Logis-LCA and Spline-LCA outperformed competing methods for the LCA model and Logis-LCA outperformed PCA+Infomax for the noisy-ICA model. These results suggest that Logis-LCA

Figure 5: Selected brain networks estimated from the HCP ToM data using Spline-LCA. The first row depicts a task-activated component that was highly correlated with the mentalizing (green) and random (blue) tasks (MNI coordinates: 35,-75,8; thresholded $|s_{v1}| \geq 2$); a similar component was found using PCA+ProDenICA (not depicted). The second row appears to be an artifact (MNI: 0,-50,0; unthresholded); this component was not found by PCA+ProDenICA.



and Spline-LCA can be used to reveal structure for a large class of non-Gaussian observations. In our fMRI application, we simultaneously achieve dimension reduction and latent variable extraction for large image data ($T = 272$ and $V = 230,459$) and identify an artifact not identified by PCA+ICA.

The presence of local maxima in LCA can increase computational expenses, and more initializations are required for larger values of T . Since the set of orthogonal matrices is non-convex, local optima are also a problem in PCA+ICA (e.g., Risk et al. 2014). For fMRI data, fifty initializations appeared to be adequate when estimating thirty components with nearly three hundred time points. In general, we found that Logis-LCA was less sensitive to initialization than Spline-LCA (results not shown). It appears that the additional flexibility of Spline-LCA comes at the expense of increased detection of local maxima. We favor Spline-LCA because it can estimate sub-Gaussian densities. However, sub-Gaussian components appear to be uncommon in fMRI (sparse images are super-Gaussian). Future research should examine whether Spline-LCA offers advantages over Logis-LCA in fMRI. Additionally, developing algorithms to more efficiently address local optima is an avenue for future research.

An important advantage of LCA over existing frameworks is its robustness to misspecification of the number of estimated components. This robustness suggests LCA could be used to improve estimates of functional connectivity in fMRI studies. In contrast, estimating the correct number of components in noisy ICA is a pre-requisite to recovering valid components (Section 4, see also Allasonniere and Younes 2012). Beckmann and Smith (2004) explored the use of probabilistic PCA to estimate the number of brain networks prior to ICA in order to avoid model over-fitting, which addresses the concern that over fitting may separate a single network into multiple networks. However, our simulations suggest that using too few components leads to inappropriately aggregated networks in PCA+ICA methods (Figures 2 and 3). In contrast, the components recovered for $Q^* \neq Q$ in Logis-LCA across model scenarios (Figures 2 and 3) and Spline-LCA for the LCA scenario (Figure 2) accurately represent functional connectivity. In the leaf data example, the first two components were nearly identical for $Q^* = 2$ and $Q^* = 5$ for LCA but differed for PCA+ICA. Although we have argued that our framework is robust to the choice of Q^* , we would like a rigorous method to determine the number of components. For iid data, AIC may be an effective method, but AIC and other model selection criteria are ineffective when correlation among observations is not taken into account adequately. Future research should investigate selection criteria for non-iid data.

LCA offers a computationally tractable alternative to one of the most common applications of ICA to fMRI and EEG: artifact detection. Currently, PCA+ICA is used as a

pre-processing step to reveal biologically implausible loadings and/or loadings resembling physiological artifacts that can be used to de-noise data for subsequent analyses (Beckmann, 2012). In LCA, these artifacts appear as LCs since they have non-Gaussian distributions. Our detection of the artifact in component two (Figure 5) suggests LCA could be used for more powerful denoising methods over traditional PCA+ICA. Artifacts may increase and/or become more problematic when using state-of-the-art data with high-resolution, as smaller voxels are associated with smaller signals, suggesting artifact removal is increasingly important (Griffanti et al., 2014). The HCP data represent the highest resolution and fastest acquisition times currently available in fMRI, and thus LCA offers a promising alternative to ICA for artifact detection.

8 Acknowledgments

We thank Dr. Nathan Spreng, Department of Human Development, Cornell University, for scientific guidance and assistance with the Human Connectome data. Research was supported by a Xerox PARC Faculty Research Award and NSF grant DMS-1455172. Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

A Proofs

We assume all random variables are mean zero. In Kagan et al. (1973), a random variable $\mathbf{X} \in \mathbb{R}^T$ is said to have a *linear structure* if it can be represented as $\mathbf{X} = \mathbf{B}\mathbf{Y}$ where the elements of \mathbf{Y} are mutually independent random variables and no two columns of \mathbf{B} are proportional. We say a linear-structure random vector \mathbf{X} has *essentially unique structure* if for any two representations $\mathbf{X} = \mathbf{B}\mathbf{Y}$ and $\mathbf{X} = \mathbf{C}\mathbf{Z}$, we have \mathbf{B} equals \mathbf{C} up to scaling and column permutation, which we denote as $\mathbf{B} \cong \mathbf{C}$. A random variable \mathbf{X} is non-unique if there exist representations $\mathbf{X} = \mathbf{B}\mathbf{Y} = \mathbf{C}\mathbf{Z}$ but $\mathbf{B} \not\cong \mathbf{C}$. Let $\stackrel{d}{=}$ denote equal in distribution. First consider the theorem on uniqueness of decomposition.

Theorem 10.3.9 from Kagan et al. (1973). *Let $\mathbf{X} = \mathbf{A}\mathbf{Y}$ be a structural representation of \mathbf{X} and let the columns of \mathbf{A} be linearly independent. Then \mathbf{X} can be expressed as $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$, where \mathbf{X}_1 and \mathbf{X}_2 are independent, \mathbf{X}_1 has essentially unique structure and \mathbf{X}_2 is multivariate normal with a non-unique structure. Moreover, this decomposition is unique*

in the sense that if $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$ is another decomposition, where \mathbf{Z}_1 has essentially unique structure, \mathbf{Z}_2 is multivariate normal, and \mathbf{Z}_1 is independent of \mathbf{Z}_2 , then $\mathbf{Z}_1 \stackrel{d}{=} \mathbf{X}_1$ and $\mathbf{Z}_2 \stackrel{d}{=} \mathbf{X}_2$.

For a proof see Kagan et al. (1973).

Before proving Theorem 1, we consider the following lemma.

Lemma 1. *Suppose \mathbf{Z} and \mathbf{X} each have essentially unique structure and $\mathbf{Z} \stackrel{d}{=} \mathbf{X}$. Consider their structural representations: $\mathbf{Z} = \mathbf{M}_S \mathbf{S}$ and $\mathbf{X} = \mathbf{M}_S^* \mathbf{S}^*$ where $\mathbf{M}_S \in \mathbb{R}^{T \times Q}$ and $\mathbf{M}_S^* \in \mathbb{R}^{T \times Q}$ for $Q \leq T$, and $\text{rank}(\mathbf{M}_S) = \text{rank}(\mathbf{M}_S^*) = Q$. Then $\mathbf{M}_S \cong \mathbf{M}_S^*$ and $\mathbf{S} \stackrel{d}{=} \mathbf{S}^*$.*

Proof. We have $\mathbf{M}_S \stackrel{d}{=} \mathbf{M}_S^* \mathbf{S}^*$. Then,

$$(\mathbf{M}_S' \mathbf{M}_S)^{-1} \mathbf{M}_S' \mathbf{M}_S \mathbf{S} = (\mathbf{M}_S' \mathbf{M}_S)^{-1} \mathbf{M}_S' \mathbf{M}_S^* \mathbf{S}^*.$$

Letting $\mathbf{B} = (\mathbf{M}_S' \mathbf{M}_S)^{-1} \mathbf{M}_S' \mathbf{M}_S^*$, we have $\mathbf{S} \stackrel{d}{=} \mathbf{B} \mathbf{S}^*$. Now \mathbf{S} has non-Gaussian independent components and thus has essentially unique structure (Theorem 10.3.5 in Kagan et al. 1973); in particular, $\mathbf{S} = \mathbf{I} \mathbf{S}$. We can define a random variable $\mathbf{R} = \mathbf{B}^{-1} \mathbf{S}$, and note that $\mathbf{R} \stackrel{d}{=} \mathbf{S}^*$, and \mathbf{S}^* has independent components, which implies \mathbf{R} has independent components, which implies $\mathbf{B} \mathbf{R}$ is a structural representation of \mathbf{S} . Since \mathbf{S} has essentially unique structure, $\mathbf{B} \cong \mathbf{I}$. It follows that $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations.

Without loss of generality we can scale \mathbf{S} and \mathbf{S}^* such that $\mathbf{E} \mathbf{S} \mathbf{S}' = \mathbf{E} \mathbf{S}^* \mathbf{S}^{*'} = \mathbf{I}$. Now we have $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ and $\mathbf{M}_S^* \mathbf{S}^* \stackrel{d}{=} \mathbf{M}_S \mathbf{S}$. Towards a contradiction, suppose $\mathbf{M}_S^* \not\cong \mathbf{M}_S$. But then $\mathbf{M}_S^* \mathbf{S}^* \stackrel{d}{\neq} \mathbf{M}_S \mathbf{S}$. \square

We now prove Theorem 1.

Theorem 1. *Let $\mathbf{X} = \mathbf{M}_S \mathbf{S} + \mathbf{M}_N \mathbf{N}$, where $\mathbf{M}_S \in \mathbb{R}^{T \times Q}$, the elements of \mathbf{S} are mutually independent non-Gaussian components, $\mathbf{M}_N \in \mathbb{R}^{T \times (T-Q)}$, \mathbf{N} is $(T-Q)$ -variate normal, and $[\mathbf{M}_S, \mathbf{M}_N]$ is full rank. Then for any other representation $\mathbf{X} = \mathbf{M}_S^* \mathbf{S}^* + \mathbf{E}^*$ where $\mathbf{S}^* \in \mathbb{R}^Q$ are independent non-Gaussian components and \mathbf{E}^* is multivariate normal, we have: $\mathbf{M}_S^* \cong \mathbf{M}_S$; $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations; and $\mathbf{E}^* \stackrel{d}{=} \mathbf{M}_N \mathbf{N}$.*

Proof. Since \mathbf{X} has a unique decomposition in the sense of Theorem 10.3.9, we have $\mathbf{M}_S \mathbf{S} \stackrel{d}{=} \mathbf{M}_S^* \mathbf{S}^*$ and $\mathbf{M}_N \mathbf{N} \stackrel{d}{=} \mathbf{E}^*$. Moreover, $\mathbf{M}_S \mathbf{S}$ and $\mathbf{M}_S^* \mathbf{S}^*$ have essentially unique structure (Theorem 10.3.5 in Kagan et al. 1973). Applying Lemma 1, we obtain the desired result. \square

Corollary 1. *Suppose the linear structure model in (1) of the main manuscript with density defined in (2) and suppose that Assumption 1 holds. Then $\{f_1, \mathbf{w}_1\}, \dots, \{f_Q, \mathbf{w}_Q\}$ are identifiable up to sign and ordering. Note the rows \mathbf{w}_{Q+k} for $k = 1, \dots, T-Q$ are not identifiable.*

Proof. For identifiability, we need to show that if there exist densities g_1, \dots, g_T and a matrix \mathbf{C} such that

$$|\det(\mathbf{L})| \prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{L} \mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L} \mathbf{x}) = |\det(\mathbf{C})| \prod_{\ell=1}^T g_\ell(\mathbf{c}'_\ell \mathbf{x}) \quad (12)$$

then Q of the marginal densities g_1, \dots, g_T are equivalent to f_1, \dots, f_Q and the rest of the distributions are Gaussian, and that each of the corresponding Q rows of \mathbf{C} equal $\mathbf{w}'_1 \mathbf{L}, \dots, \mathbf{w}'_Q \mathbf{L}$. Using a change of variable $\mathbf{Z} = \mathbf{L} \mathbf{x}$, we consider the model $\mathbf{Z} = \mathbf{A}_S \mathbf{S} + \mathbf{A}_N \mathbf{N}$, such that $[\mathbf{w}'_1; \dots; \mathbf{w}'_Q] = \mathbf{A}'_S$ (where $[\mathbf{w}'_1; \dots; \mathbf{w}'_Q]$ indicates stacked row vectors) and $[\mathbf{w}'_{Q+1}; \dots; \mathbf{w}'_T] = \mathbf{A}'_N$. Then (12) is equivalent to

$$\prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{z}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{z}) = |\det(\mathbf{C})| |\det(\mathbf{L})|^{-1} \prod_{\ell=1}^T g_\ell(\mathbf{c}'_\ell \mathbf{L} \mathbf{z}).$$

We define $\mathbf{R} = \mathbf{C} \mathbf{L}^{-1}$ such that we have

$$\prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{z}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{z}) = |\det(\mathbf{R})| \prod_{\ell=1}^T g_\ell(\mathbf{r}'_\ell \mathbf{z}). \quad (13)$$

We have demonstrated identifiability up to signed permutations if we can show that Q of the marginal densities g_1, \dots, g_T are equivalent to f_1, \dots, f_Q ; that each of the corresponding Q rows of \mathbf{R} equal $\pm \mathbf{w}_1, \dots, \pm \mathbf{w}_Q$; and that $|\det(\mathbf{R})| = 1$.

Define $\mathbf{K} = \mathbf{R}^{-1}$. Given the relationship in (13), then there exists another *linear structure* representation of \mathbf{Z} such that $\mathbf{Z} = \mathbf{K} \mathbf{Y}$. Without loss of generality, we have $\mathbf{E} \mathbf{Y} \mathbf{Y}' = \mathbf{I}$ (there is no loss of generality because we can scale \mathbf{K} such that $\mathbf{E} \mathbf{Y} \mathbf{Y}' = \mathbf{I}$). From Theorem 10.3.3 in Kagan et al. (1973), \mathbf{Z} has the decomposition $\mathbf{Z} = \mathbf{K}_1 \mathbf{Y}_1 + \mathbf{K}_2 \mathbf{Y}_2$ in which \mathbf{Y}_1 are independent non-Gaussian and \mathbf{Y}_2 are Gaussian. Then from Theorem 1 and the assumption of unit variance, we have that $\mathbf{Y}_1 \stackrel{d}{=} \mathbf{S}$ (up to ordering), and it follows that there exists a subset of g_1, \dots, g_T equal to f_1, \dots, f_Q . Also from Theorem 1, we have $\mathbf{K}_1 \cong \mathbf{A}_S$. Note that $\mathbf{K} \in \mathcal{O}_{T \times T}$ since $\mathbf{E} \mathbf{Y} \mathbf{Y}' = \mathbf{I}$ and $\mathbf{E} \mathbf{Z} \mathbf{Z}' = \mathbf{I}$. Then the scaling of \mathbf{K}_1 is also identifiable such that there exists a signed permutation matrix, \mathbf{P}_\pm , such that $\mathbf{K}_1 \mathbf{P}_\pm = \mathbf{A}_S$. Note that $\mathbf{W}_S = \mathbf{A}'_S$. Define $\mathbf{R}_S = \mathbf{K}'_1$. Then $\mathbf{P}'_\pm \mathbf{R}_S = \mathbf{W}_S$. \square

Proposition 3. Consider a random vector $\mathbf{Y} \in \mathbb{R}^T$ with density $f_{\mathbf{Y}}$ such that $\mathbf{E} \mathbf{Y} = \mathbf{0}$ and $\mathbf{E} \mathbf{Y} \mathbf{Y}' = \mathbf{I}_T$. Then for any \mathbf{o} and \mathbf{w} such that $\mathbf{o}' \mathbf{o} = \mathbf{w}' \mathbf{w} = 1$, we have

$$\mathbf{E} \log \phi(\mathbf{o}' \mathbf{Y}) = \mathbf{E} \log \phi(\mathbf{w}' \mathbf{Y}).$$

Proof. We can ignore the normalizing constants of $\phi(x)$ and consider the quadratic term of the Gaussian kernel. Then we have $E(\mathbf{o}'\mathbf{Y})^2 = \mathbf{o}'E(\mathbf{Y}\mathbf{Y}')\mathbf{o} = \mathbf{o}'\mathbf{I}\mathbf{o} = \mathbf{o}'\mathbf{o} = 1$ and similarly for $E(\mathbf{w}'\mathbf{Y})^2$. \square

Next, we show the consistency of the oracle estimate of \mathbf{W}_S .

Theorem 2. *Suppose \mathbf{X} follows the LCA model in (1) of the main manuscript with Assumption 1 holding and assume the non-Gaussian components have bounded absolutely continuous densities (satisfied by the classes considered below). Additionally assume $E\mathbf{X} = \mathbf{0}$ and $E\mathbf{X}\mathbf{X}' = \mathbf{I}$ (here, \mathbf{W}_S comprises the first Q rows of \mathbf{M}^{-1}). Given an iid sample $\mathbf{x}_1, \dots, \mathbf{x}_V$, $\widehat{\mathbf{W}}_S^{Tr} \rightarrow \mathbf{W}_S$ almost surely on the equivalence class of signed permutations.*

Proof. Note that $\mathcal{O}_{Q \times T}$ is compact. We will show the four assumptions in Wald's consistency proof as recast in Pollard (2001) are satisfied. Let f_S denote the joint density of the LCs. First, we show $E \log f_S(\mathbf{O}_S \mathbf{X}) \leq E \log f_S(\mathbf{W}_S \mathbf{X})$ for any $\mathbf{O}_S \in \mathcal{O}_{Q \times T}$ with equality if and only if $\mathbf{O}_S \cong \mathbf{W}_S$. Let \mathbf{W}_N denote rows $T - Q + 1$ to T of \mathbf{W} . (That $E \log f_S(\mathbf{O}_S \mathbf{X}) \leq E \log f_S(\mathbf{W}_S \mathbf{X})$ does not hold trivially can be seen by the following argument:

$$\begin{aligned} E \log \frac{f_S(\mathbf{O}_S \mathbf{X})}{f_S(\mathbf{W}_S \mathbf{X})} &\leq \log E \frac{f_S(\mathbf{O}_S \mathbf{X})}{f_S(\mathbf{W}_S \mathbf{X})} \\ &= \log \int \left\{ \frac{f_S(\mathbf{O}_S \mathbf{x})}{f_S(\mathbf{W}_S \mathbf{x})} \right\} \{f_S(\mathbf{W}_S \mathbf{x}) \phi(\mathbf{W}_N \mathbf{x})\} d\mathbf{x} \\ &= \log \int f_S(\mathbf{O}_S \mathbf{x}) \phi(\mathbf{W}_N \mathbf{x}) d\mathbf{x}. \end{aligned}$$

We would like the last quantity to be equal to zero, in which case we would obtain the desired bound. Let \mathbf{W}^* be the $T \times T$ matrix formed by stacking \mathbf{O}_S and \mathbf{W}_N . The term $f_S(\mathbf{O}_S \mathbf{x}) \phi(\mathbf{W}_N \mathbf{x})$ is a density if and only if $|\det(\mathbf{W}^*)| = 1$, which is not true in general because \mathbf{O}_S may not be orthogonal to \mathbf{W}_N . Consequently, this quantity could integrate to greater than one, in which case we would have $E \log f_S(\mathbf{O}_S \mathbf{X}) \leq E \log f_S(\mathbf{W}_S \mathbf{X}) + \alpha$ for some $\alpha > 0$, and thus our bound is not tight enough.)

Define an orthogonal matrix in $\mathcal{O}_{T \times T}$ such that rows 1 to Q are equal to \mathbf{O}_S and the other rows are arbitrary. Then

$$\begin{aligned} E \log \frac{f_S(\mathbf{O}_S \mathbf{X})}{f_S(\mathbf{W}_S \mathbf{X})} &= E \log \frac{f_S(\mathbf{O}_S \mathbf{X}) \phi(\mathbf{O}_N \mathbf{X})}{f_S(\mathbf{W}_S \mathbf{X}) \phi(\mathbf{O}_N \mathbf{X})} \\ &= E \log \frac{f_S(\mathbf{O}_S \mathbf{X}) \phi(\mathbf{O}_N \mathbf{X})}{f_S(\mathbf{W}_S \mathbf{X}) \phi(\mathbf{W}_N \mathbf{X})}, \end{aligned}$$

where the second line follows from Proposition 1. Then

$$\begin{aligned} \mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{X})} &\leq \log \mathbb{E} \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{X})} \\ &= \log \int f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{x})d\mathbf{x} \\ &= 0, \end{aligned}$$

which holds with equality if and only if $f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{x}) = f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x})$, where the only if direction is a consequence of absolute continuity. Now suppose equality holds for the matrix $\mathbf{O}_{\mathbf{S}}^*$ and let \mathbf{Y} be a random variable with density $f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}^*\mathbf{y})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{y}) = f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{y})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{y})$. Let $\mathbf{O}_+ = [\mathbf{O}_{\mathbf{S}}^*, \mathbf{O}_{\mathbf{N}}']'$. Then there exist random variables \mathbf{R}_+ and \mathbf{R} such that $\mathbf{Y} = \mathbf{O}_+\mathbf{R}_+$ and $\mathbf{Y} = \mathbf{W}\mathbf{R}$. Applying Theorem 1, we have $\mathbf{O}_{\mathbf{S}}^* \cong \mathbf{W}_{\mathbf{S}}$. It follows that

$$\mathbb{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X}) < \mathbb{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$$

for all $\mathbf{O}_{\mathbf{S}} \not\cong \mathbf{W}_{\mathbf{S}}$. The other three conditions are satisfied since we assume continuous, bounded densities and our estimator is an M-estimator. \square

Proposition 4. *Let G be the class of all cubic splines $g : \mathbb{R} \rightarrow \mathbb{R}$. Consider the argmax of (8) of the main manuscript for $g_q \in G$ with g_q denoting the tilt function for the q th component. Then (i) $\int \phi(x)e^{g_q(x)} dx = 1$ and (ii) $\int x\phi(x)e^{g_q(x)} dx = 0$ for each q .*

Proof. It suffices to consider the case $Q^* = 1$. Let \mathbf{o}_1 be given and define $s_v = \mathbf{o}_1' \widehat{\mathbf{L}}\mathbf{x}_v$. Define the class of functions $H = \{h : \mathbb{R} \rightarrow \mathbb{R}, h(x) = \theta_0 + \theta_1 x, \theta_0, \theta_1 \in \mathbb{R}\}$, and note that H is in the null space of the penalty $\lambda_1 \int \{g_1''(x)\}^2 dx$. Let $J = \{j \in G : \langle h, j \rangle = 0 \forall h \in H\}$. Then $G = H \oplus J$, where \oplus denotes the direct sum. Now let g^* be the argmax of (8) of the main manuscript for $g^* \in G$. Then $g^*(x) = h^*(x) + j^*(x)$ for some $h^* \in H$, $j^* \in J$. Then, with $\ell(\cdot)$ the penalized log-likelihood given by (8) of the main paper, we have

$$\frac{\partial \ell(g^*)}{\partial \theta_0} = 1 - \int \phi(x)e^{g^*(x)} dx,$$

from which it follows that $\phi(x)e^{g^*(x)}$ is a density. Next,

$$\frac{\partial \ell(g^*)}{\partial \theta_1} = \frac{1}{V} \sum_{v=1}^V s_v - \int x\phi(x)e^{g^*(x)} dx,$$

where we have applied Leibnitz's rule to interchange differentiation with respect to θ_1 and integration with respect to x since $\phi(x)e^{g^*(x)}$ and $x\phi(x)e^{g^*(x)}$ are continuous on \mathbb{R} . Then

it follows that $\mathbb{E} S = 0$ for S with density $\phi(x)e^{g^*(x)}$. \square

B Additional Background

B.1 Projection Pursuit, D-FastICA, and Non-Gaussian Component Analysis

Projection pursuit is an exploratory method for finding low-dimensional representations of multivariate data that reveal interesting patterns and structure (Huber, 1985). Let \mathbf{x}_v , $v = 1, \dots, V$ be a data sample with $\mathbf{x}_v \in \mathbb{R}^T$, and assume $\sum_{v=1}^V \mathbf{x}_v = \mathbf{0}$, where $\mathbf{0}$ is the vector of T zeros, and $\frac{1}{V} \sum_{v=1}^V \mathbf{x}_v^2 = \mathbf{1}$, where $\mathbf{1}$ is a length T vector of ones. Let Q be the number of projection pursuit directions that are estimated. In FastICA in deflation mode (D-FastICA), the projection pursuit index is equivalent to an approximation of negentropy (Hyvarinen, 1999):

$$\mathbf{w}_q = \underset{\mathbf{w} \in \mathbb{R}^T}{\operatorname{argmax}} \left\{ \frac{1}{V} \sum_{v=1}^V G(\mathbf{w}'\mathbf{x}_v) - \mathbb{E} G(n) \right\}^2, \quad (14)$$

where \mathbf{w} is orthogonal to $\widehat{\mathbf{w}}_1, \dots, \widehat{\mathbf{w}}_{q-1}$ and $\|\mathbf{w}\| = 1$ with $\|\cdot\|$ denoting the L2-norm, G is a non-linear function, and n is a standard normal random variable. A common choice for G is $\log \cosh(x)$, which will be used to estimate projection pursuit directions in our simulations.

NGCA uses multiple projection pursuit indices (Blanchard et al., 2006) or radial basis functions (Kawanabe et al., 2007) to find a non-Gaussian subspace that is assumed to contain the interesting features of data. NGCA can be formulated using a semiparametric likelihood,

$$f_{\mathbf{X}}(\mathbf{x}) = h(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi_{\mathbf{0},\mathbf{\Sigma}}(\mathbf{x}) \quad (15)$$

where $\phi_{\mathbf{0},\mathbf{\Sigma}}$ is multivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$; $\mathbf{W}_{\mathbf{S}}$ is a $Q \times T$ matrix; and $h(\cdot)$ is a function that captures departures from Gaussianity under the constraint that $f_{\mathbf{X}}(\mathbf{x})$ is a density. The NGCA model does not assume independent factors, and thus we do not consider it in our simulations.

The density in the Spline-LCA model can be considered an extension of (15) with the additional assumption of independence.

Proposition 5. *Let \mathbf{X} be a random variable from the LCA model where the LCs have tilted*

Gaussian densities. Then the density of \mathbf{X} is

$$f_{\mathbf{X}}(\mathbf{x}) = \phi_{\mathbf{0}, \Sigma}(\mathbf{x}) \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})}$$

where $\phi_{\mathbf{0}, \Sigma}$ is the mean zero multivariate distribution with covariance Σ .

Proof. Using the tilted Gaussian density, we have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \det \mathbf{L} \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})} \phi(\mathbf{w}'_q \mathbf{L} \mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L}' \mathbf{x}) \\ &= \left\{ \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})} \right\} (2\pi)^{-T/2} (\det \mathbf{L}) \exp \left\{ -\frac{1}{2} \sum_{k=1}^T \mathbf{x}' \mathbf{L} \mathbf{w}_k \mathbf{w}'_k \mathbf{L} \mathbf{x} \right\} \\ &= (\det \Sigma)^{-1/2} (2\pi)^{-T/2} \exp \left\{ -\frac{1}{2} \mathbf{x}' \Sigma^{-1} \mathbf{x} \right\} \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})}. \end{aligned}$$

□

Writing the likelihood in this way, one notes that we are using the Gaussian density to model the covariance between components and we are using the tilt functions to model deviations from the Gaussian model.

B.2 Noisy ICA and IFA

In the noisy ICA model, Q ICs are corrupted by rank- T Gaussian noise, where $Q \leq T$ (Hyvärinen et al., 2001),

$$\mathbf{X} = \mathbf{M}_S \mathbf{S} + \mathbf{E} \tag{16}$$

with $\mathbf{X} \in \mathbb{R}^T$, \mathbf{M}_S is $T \times Q$ with $Q \leq T$, \mathbf{E} is mean-zero multivariate normal with covariance matrix Ψ , and \mathbf{E} is independent of \mathbf{S} .

Assume that $\Psi = \sigma^2 \mathbf{I}$. Let $\lambda_1, \dots, \lambda_Q$ denote the eigenvalues from the covariance matrix of $\mathbf{M}_S \mathbf{S}$ and let $\lambda_{\epsilon_1}, \dots, \lambda_{\epsilon_T}$ denote the eigenvalues from the decomposition of \mathbf{E} . Under the assumption of isotropic noise, we have $\lambda_{\epsilon_i} = \lambda_{\epsilon_j} = \sigma^2$ for all $i, j = 1, \dots, T$. Then the eigenvalue decomposition can be written as

$$\text{Cov } \mathbf{X} = \mathbf{U} \text{diag}(\lambda_1 + \sigma^2, \dots, \lambda_Q + \sigma^2, \sigma^2, \dots, \sigma^2) \mathbf{U}'.$$

Let \mathbf{X}_{data} be the $V \times T$ data matrix. In PCA+ICA, noise-free ICA is applied to the first Q left singular vectors of \mathbf{X}_{data} multiplied by \sqrt{V} , which is equivalent to the first Q standardized principal components.

In IFA, (16) is estimated under the assumption that the densities of the ICs are Gaussian mixtures (Attias, 1999). In its original formulation, Ψ was an arbitrary positive definite matrix, the IC densities had K_q classes, and the variance of each IC was standardized to unity after each iteration. In our presentation and estimation, we assume that the covariance of the noise is $\sigma^2 \mathbf{I}$ and IC densities are mixtures of two Gaussians, which has been assumed elsewhere (e.g., Guo and Tang 2013; Beckmann and Smith 2004), and enforce the constraint that the IC densities are mean zero with unit variance. Let π_{q1} be the probability that an observation of the q th IC comes from the first class, where the first class has a normal distribution with mean μ_{q1} and variance ν_{q1} . Then the probability, mean, and variance for the second class are $\pi_{q2} = 1 - \pi_{q1}$, $\mu_{q2} = -\frac{\pi_{q1}\mu_{q1}}{\pi_{q2}}$, and $\nu_{q2} = \frac{1 - \pi_{q1}\nu_{q1} - \pi_{q1}\mu_{q1}^2}{\pi_{q2}} - \mu_{q2}^2$, respectively. Then the joint density of \mathbf{x}_v can be written

$$f_{\mathbf{x}}(\mathbf{x} \mid \mathbf{M}_{\mathbf{S}}) = \prod_{t=1}^T \int \phi_{0, \sigma^2}(\mathbf{x}_t - \mathbf{m}'_t \mathbf{s}) f_{\mathbf{s}}(\mathbf{s}) d\mathbf{s}, \quad (17)$$

where ϕ_{0, σ^2} is a normal density with mean zero and variance σ^2 and

$$f_{\mathbf{s}}(\mathbf{s}) = \prod_{q=1}^Q \left\{ \pi_{q1} \phi_{\mu_{q1}, \nu_{q1}}(s_q) + \pi_{q2} \phi_{\mu_{q2}, \nu_{q2}}(s_q) \right\}.$$

Analytic integration across \mathbf{s} is possible. Let k_q equal one if s_q is in the first class and zero otherwise. Let \mathcal{K} be the set of all possible states for the Q components composed from the Cartesian product Q -times of the singletons $\{\{0\}, \{1\}\}$. Let $\mathbf{k}_j = \{k_1, \dots, k_Q\}$ denote an element of \mathcal{K} , where $j \in \{1, \dots, 2^Q\}$. Let $\boldsymbol{\mu}(\mathbf{k}_j)$ and $\boldsymbol{\nu}(\mathbf{k}_j)$ denote the conditional means of \mathbf{s} given the states \mathbf{k}_j . Now define

$$\Sigma(\mathbf{k}_j) = \mathbf{M}_{\mathbf{S}} \text{diag}\{\boldsymbol{\nu}(\mathbf{k}_j)\} \mathbf{M}'_{\mathbf{S}} + \sigma^2 \mathbf{I}$$

and

$$\boldsymbol{\mu}^*(\mathbf{k}_j) = \mathbf{M}_{\mathbf{S}} \boldsymbol{\mu}(\mathbf{k}_j).$$

Then the density is

$$f_{\mathbf{x}}(\mathbf{x} \mid \mathbf{M}_{\mathbf{S}}) = \sum_{\mathbf{k}_j \in \mathcal{K}} \Phi\{\mathbf{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \Sigma(\mathbf{k}_j)\} \prod_{q=1}^Q \pi_{q1}^{k_q} \pi_{q2}^{1-k_q} \quad (18)$$

with $\Phi\{\mathbf{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\mathbf{k}_j)\}$ multivariate normal with mean $\boldsymbol{\mu}^*(\mathbf{k}_j)$ and variance $\boldsymbol{\Sigma}(\mathbf{k}_j)$ (see (16) and (17) in Attias 1999). Then a likelihood can be constructed from (18), and given some $\widehat{\mathbf{M}}_{\mathbf{S}}$, the ICs can be estimated from their conditional means. Alternatively, maximum a posteriori estimates of the ICs could be obtained, though we pursue the former here.

C Using the fixed-point algorithm to fit the LCA model

Here we describe the fixed-point algorithm from Hyvarinen (1999). Our account is equivalent to Hyvarinen (1999) except we use our novel discrepancy measure (*PMSE*) and a different orthogonalization method. Under the constraint that the noise components follow a standard normal distribution, we can ignore rows $Q^* + 1 : T$ in $\widehat{\mathbf{W}}$. For now, we assume the densities of the latent components f_1, \dots, f_{Q^*} , are known. Define the scalar $h_q(x) = \log f_q(x)$, and let $h'(x)$ denote its derivative. Algorithm 1 provides details on estimating $\widehat{\mathbf{W}}_{\mathbf{S}}$.

Algorithm 2: The fastICA algorithm for LCA.

Inputs : The whitened $V \times T$ data matrix \mathbf{Z} ; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ .

Result: Estimates of the latent components, $\widehat{\mathbf{S}} = \mathbf{Z}\widehat{\mathbf{W}}'_{\mathbf{S}}$.

1. Let $\mathbf{S}^{(0)} = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{(0)'}$ and let $n = 0$.
2. For each $q = 1, \dots, Q$, calculate

$$\mathbf{w}_q^* = \frac{1}{V} \sum_{v=1}^V \{ \mathbf{z}_v h'_q(\mathbf{w}_q^{(n)'} \mathbf{z}_v) - h''_q(\mathbf{w}_q^{(n)'} \mathbf{z}_v) \mathbf{w}_q^{(n)} \}$$

3. Calculate the SVD of $\mathbf{W}_{\mathbf{S}}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*'}$.
 4. Let $\mathbf{W}^{(n+1)} = \mathbf{U}^* \mathbf{V}^{*'}$.
 5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(n+1)'}, \mathbf{W}_{\mathbf{S}}^{(n)'}) < \epsilon$, stop, else increment n and repeat (2)-(3).
-

D Supplemental materials for simulations examining distributional and noise-rank assumptions

We fit D-FastICA using the ‘deflation’ option in the fastICA R package (Marchini et al., 2010). However, this popular function does not include an option to use projection pursuit for dimension reduction. If one specifies some $Q < T$ number of components, PCA is

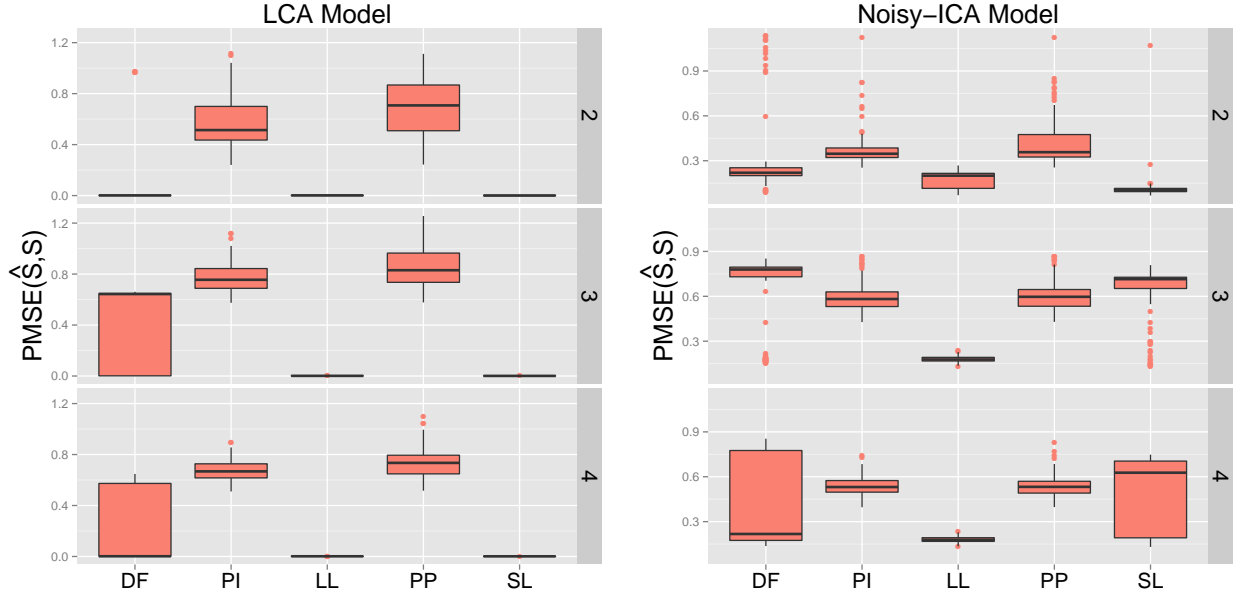
performed prior to the ICA. Consequently, one must estimate all T directions and then subset to the first two.

We fit the IFA model with two-class mixtures of normals by maximizing the log likelihood using a numerical optimizer. This contrasts with methods using approximating EM algorithms, as described in the introduction. Our implementation is not scalable to large Q or T (nor is the exact EM algorithm) but suffices for the simulation experiments. For IFA, one must specify initial values for the unmixing matrix, the variance of the isotropic noise, and the parameters of the Gaussian mixtures. We had four strategies to find the argmax as detailed here. In our function, we constrain the latent component distributions to have zero expectation and unit norm, and as a result, the number of parameters to estimate for each latent component distribution is three. First, we estimated the parameters of the model proposed in Beckmann and Smith (2004) (BS-PICA) and used this solution to initialize the IFA. We then estimated the model from six additional random matrices but with density parameters initialized from the BS-PICA solution. Secondly, when the IFA model was true, we initialized it from the true mixing matrix and true density parameters and also from six additional random matrices with density parameters initialized from their true values. When the IFA model was not true, we initialized it from the true mixing matrix but with the density parameters initialized from their BS-PICA estimates and an additional six random matrices. Thirdly, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters $(0.7, 0.7, -0.5, -0.5, 0.5, 0.5)$ (super-Gaussian distribution) for $\pi_{11}, \pi_{21}, \mu_{11}, \mu_{21}, \nu_{11}, \nu_{21}$ and $\sigma^2 = 1$. Finally, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters $(0.3, 0.3, -1, -1, 0.5, 0.5)$ (sub-Gaussian distribution) with $\sigma^2 = 1$.

The matrices \mathbf{M}_S and \mathbf{M}_N were generated by first simulating a 5×5 matrix with standard normal entries, taking the singular value decomposition (SVD), then creating a diagonal matrix with five singular values from a uniform(1,10) distribution, followed by multiplying the left singular vectors from the SVD, the diagonal matrix, and the right singular vectors, which created $[\mathbf{M}_S, \mathbf{M}_N]$. For the noisy ICA model, we generated a random mixing matrix in the same manner, then retained the first two columns.

To generate semi-orthogonal random matrices to initiate the fixed point algorithm, matrices were generated by taking the left eigenvectors from the SVD of a 2×5 matrix with entries simulated from a standard normal. We generated random matrices constrained to the principal subspace in the following manner. Let $\hat{\mathbf{U}}_{1:Q}$ denote the first Q rows from $\hat{\mathbf{U}}$ in the decomposition $\hat{\Sigma} = \hat{\mathbf{U}}\mathbf{\Lambda}\hat{\mathbf{U}}'$. Then constraining the initial matrix, $\mathbf{W}_S^{(0)}$, to the principal subspace is equivalent to $\mathbf{W}_S^{(0)} = \hat{\mathbf{U}}_{1:Q}\mathbf{O}$ where \mathbf{O} is a random $Q \times Q$ orthogonal matrix.

Figure A1: Boxplots of $PMSE$ for estimated columns of \mathbf{S} from simulations of spatial networks with temporal dependence and $Q = 3$ with $Q^* = 2, 3$, or 4. ‘DF’ = D-FastICA; ‘PI’ = PCA+Infomax; ‘LL’= Logis-LCA; ‘PP’ = PCA+ProDenICA; ‘SL’ = Spline-LCA.



E Supplemental figures for the spatio-temporal network simulations

The permutation-invariant mean squared errors for the components estimated from the spatio-temporal network simulations are much lower for Logis-LCA and Spline-LCA when the noise rank is $T - Q$ (Figure A1). When the noise is rank- T , Logis-LCA performs best. Spline-LCA is excellent at finding two of the three components, but appears to sometimes find spurious components that were produced from the correlated noise when three or four components are estimated.

F Supplemental materials for correlated multivariate data

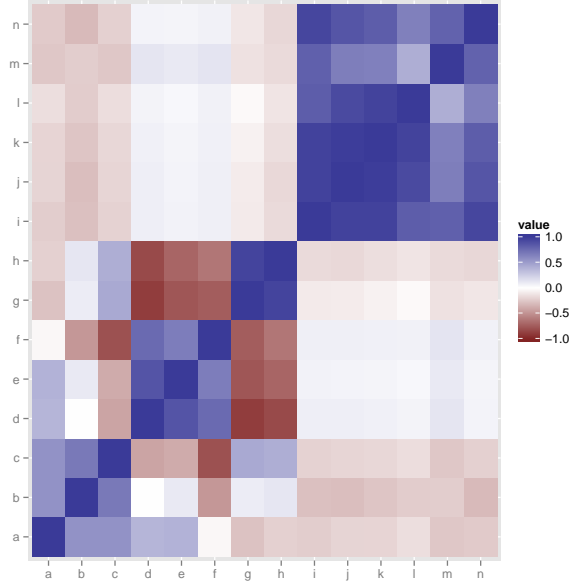
Covariates were generated from photographs of leaf samples from thirty species (Figure A2). Many of these covariates are highly correlated (Figure A3).

Logis-LCA and Spline-LCA reveal features in the data (Figures A4, A5), while PCA+Infomax and PCA+ProDenICA simply rotate the principal components. Additionally, when five components are estimated using the LCA methods, the first two components are nearly equivalent

Figure A2: Species 1-15 and 22-36 are included in the leaf dataset. Species 8 corresponds to *Neurium oleander* (blue dots in Figure 4 and Supplemental Figures 4 and 5); species 31 and 34 correspond to *Podocarpus sp.* and *Pseudosasa japonica* (green dots in Figure 4 and Supplemental Figures 4 and 5).



Figure A3: Correlation matrix of the variables in the leaf dataset: a) eccentricity, b) aspect ratio, c) elongation, d) solidity, e) stochastic convexity, f) isoperimetric factor, g) maximal indentation depth, h) lobedness, i) average intensity, j) average contrast, k) smoothness, l) third moment, m) uniformity, and n) entropy.



to the components obtained from $Q^* = 2$. This is not the case with the PCA+ICA methods. Thus, the components in LCA appear less sensitive to the number of estimated components than the components from PCA+ICA methods.

G Supplemental materials for the fMRI analysis

Whole-brain data were acquired from two sessions with 274 volumes each using gradient-echo EPI with an eight-band multifactor approach and $2 \times 2 \times 2$ mm voxels (repetition time (TR) = 720 ms; echo time (TE) = 33.1 ms; flip angle = 52° ; field of view = 208×180 mm (readout x phase-encoding); acquisition matrix = 104×90 ; slice thickness = 2.0 mm). Only the first session was used in our analyses. Inspection revealed that the first two TRs contained BOLD signals that were much higher than other time points, suggesting inadequate equilibration time. Consequently, we removed the first two TRs. After vectorization, the voxels were standardized across time to have mean zero and unit variance.

Following Risk et al. (2014), we assessed the reliability of individual components by matching components from all other initializations to the components corresponding to the argmax using the modified Hungarian algorithm. We then created dissimilarity matrices for each component based on the MSE and visualized basins of attraction using multidimensional

Figure A4: Components in the leaf data from PCA+Infomax and Logis-LCA when two components are estimated and when five components are estimated (when five components are estimated, the first two components are plotted). The green dots correspond to *Podocarpus sp.* and *Pseudosasa japonica*; the blue dots correspond to *Neurium oleander*; the red dots correspond to all other species.

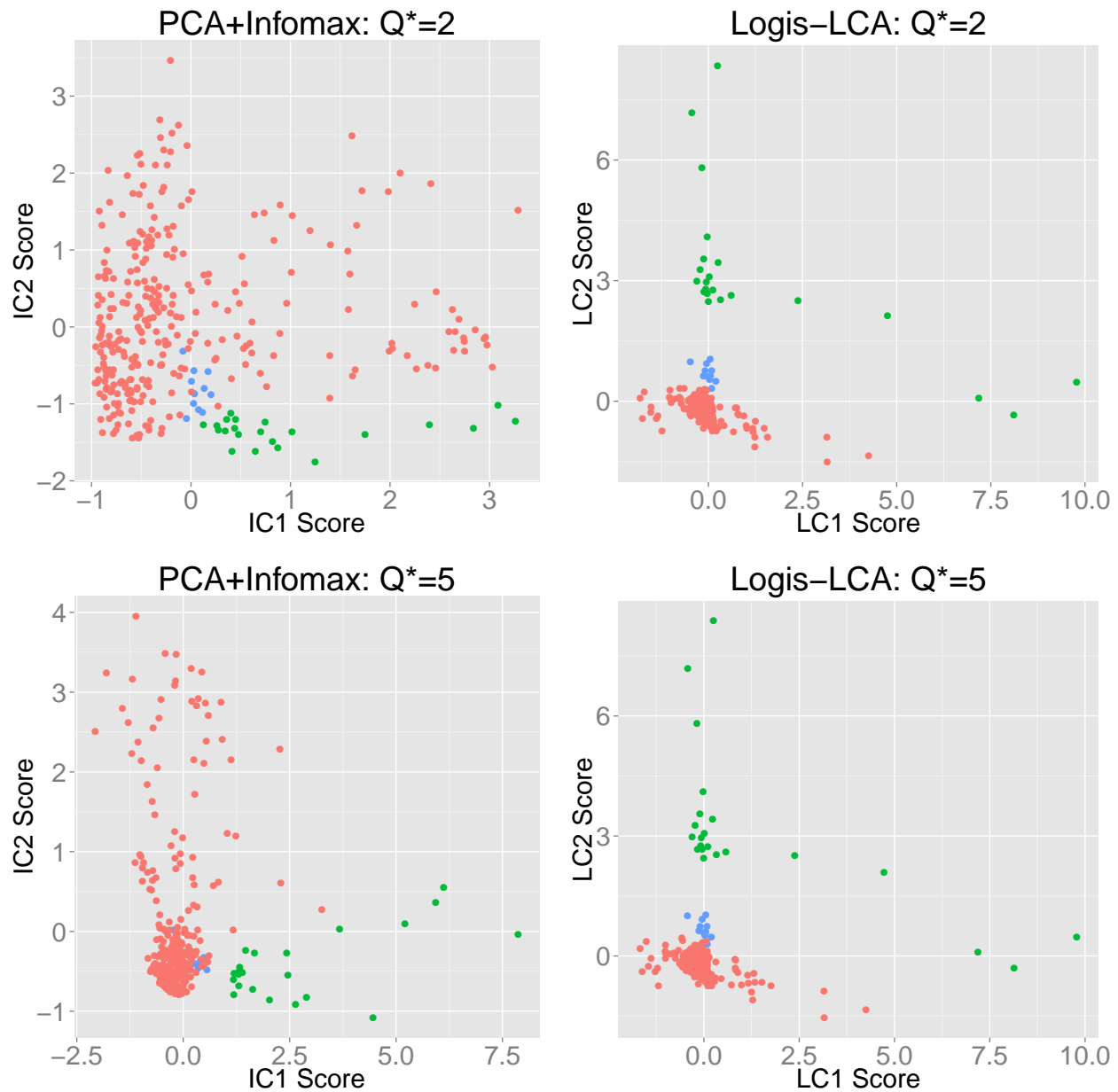
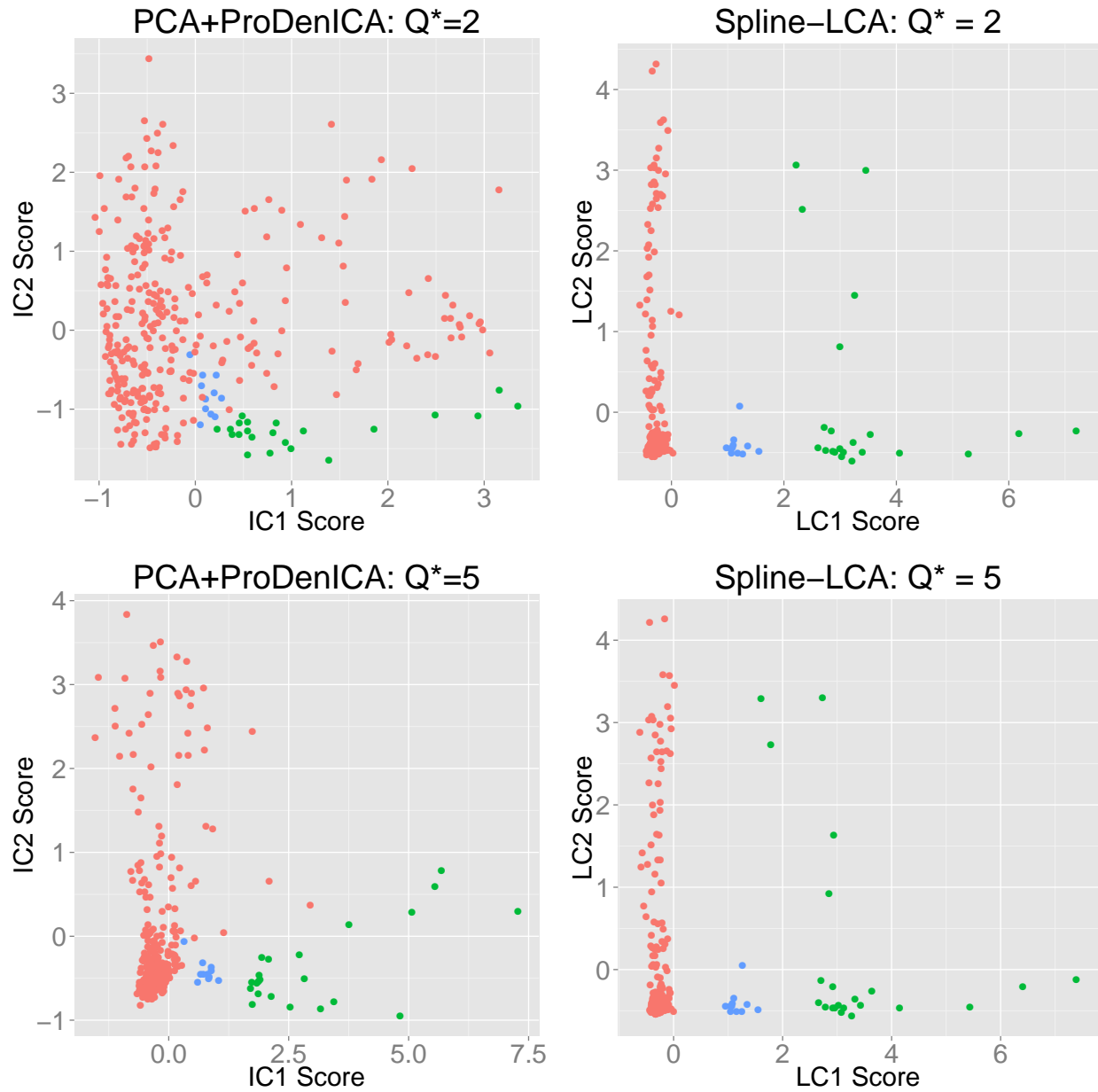


Figure A5: Components in the leaf data from PCA-ProDenICA and Spline-LCA when two components are estimated and when five components are estimated (when five components are estimated, the first two components are plotted). The green dots correspond to *Podocarpus* sp. and *Pseudosasa japonica*; the blue dots correspond to *Neurium oleander*; the red dots correspond to all other species. The plots in the first row also appear in Figure 4 of the main manuscript.



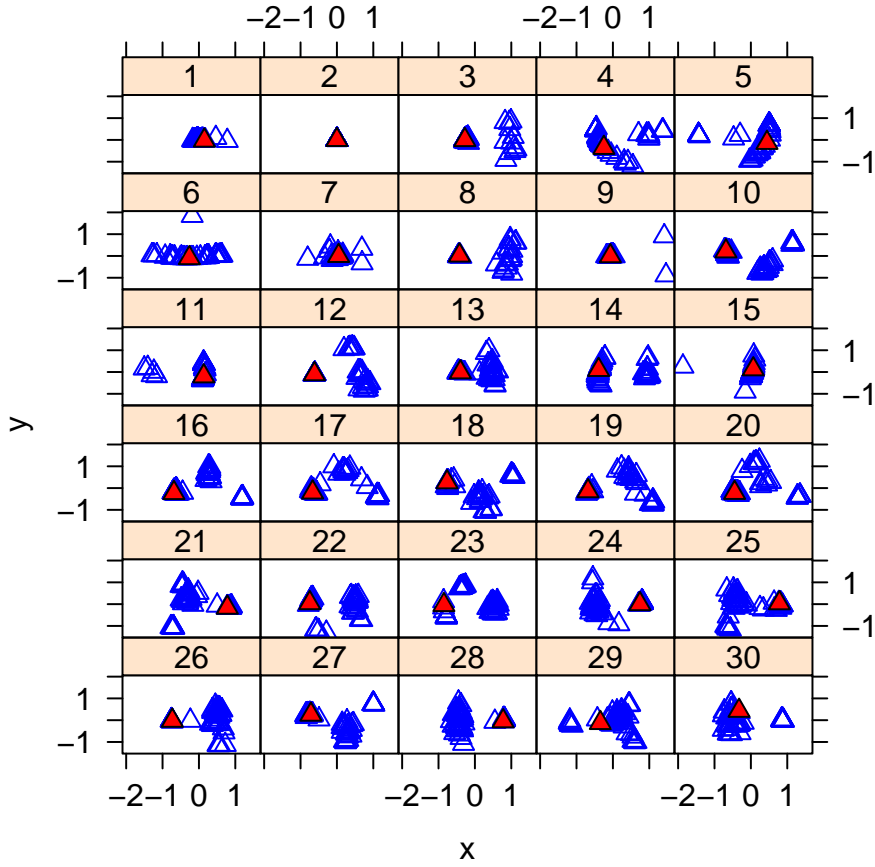


Figure A6: Multidimensional scaling of $\|\widehat{\mathbf{S}}_j^{(k)} - \widehat{\mathbf{S}}_j^{(\ell)}\|_2$ for components $j = 1, \dots, 30$ and initializations $k \neq \ell \in \{1, \dots, 30\}$. The coordinates corresponding to the initialization with the highest likelihood are depicted by solid red triangles.

mensional scaling. Generally, there were at least two basins of attraction corresponding to initializations from the principal subspace and initializations from the entire column space (Supplemental Figure A6). Components one, two, and nine were relatively robust to initialization and contained only one (main) basin of attraction. Note that in our results, we examined components one and two.

References

- Allasonniere, S. and Younes, L. (2012). A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160.
- Amari, S.-I. (1999). Natural gradient learning for over-and under-complete bases in ICA. *Neural Computation*, 11(8):1875–1883.
- Amato, U., Antoniadis, A., Samarov, A., and Tsybakov, A. (2010). Noisy independent factor analysis model for density estimation and classification. *Electronic Journal of Statistics*, 4:707–736.
- Ashburner, J., Friston, K., and Penny, W. (2004). Human brain function. *Academic press*, 1:2.
- Attias, H. (1999). Independent factor analysis. *Neural computation*, 11(4):803–851.
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189.
- Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464.
- Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, 62(2):891–901.
- Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152.
- Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.
- Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., and Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7:247–282.
- Calhoun, V. D., Liu, J., and Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45(1):S163–S172.
- Correa, N., Adali, T., and Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magnetic resonance imaging*, 25(5):684–694.
- Delorme, A., Sejnowski, T., and Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449.
- Eloyan, A. and Ghosh, S. K. (2013). A semiparametric approach to source separation using independent component analysis. *Computational Statistics and Data Analysis*, 58:383 – 396.
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124.
- Green, C. G., Nandy, R. R., and Cordes, D. (2002). PCA-preprocessing of fMRI data adversely affects the results of ICA. In *Proceedings of international society of magnetic resonance in medicine*, page 10.
- Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, 95:232–247.
- Guo, Y. and Tang, L. (2013). A hierarchical model for probabilistic independent component analysis of multi-subject fMRI studies. *Biometrics*, 69(4):970–981.
- Hastie, T. (2013). *GAM: Generalized Additive Models*. R package version 1.08.

- Hastie, T. and Tibshirani, R. (2003). Independent components analysis through product density estimation. *Advances in Neural Information Processing Systems*, 15:649–656.
- Hastie, T. and Tibshirani, R. (2010). *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*. R package version 1.0.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.
- Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, pages 435–475.
- Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.
- Hyvärinen, A., Hoyer, P., and Oja, E. (1999). Image denoising by sparse code shrinkage. In *Intelligent Signal Processing*. Citeseer.
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley-Interscience.
- Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.
- Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010). A new performance index for ICA: properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation*, pages 229–236.
- Kagan, A. M., Rao, C. R., and Linnik, Y. V. (1973). *Characterization problems in mathematical statistics*. Wiley.
- Kawanabe, M., Sugiyama, M., Blanchard, G., and Müller, K. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75.
- Lee, S., Shen, H., Truong, Y., Lewis, M., and Huang, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 106(495):1009–1024.
- Marchini, J. L., Heaton, C., and Ripley, B. D. (2010). *FastICA: FastICA Algorithms to perform ICA and Projection Pursuit*. R package version 1.1-13.
- Matteson, D. S. and Tsay, R. S. (2013). Independent component analysis via distance covariance. *ArXiv e-prints*.

- Miettinen, J., Nordhausen, K., Oja, H., and Taskinen, S. (2014). Deflation-based FastICA with adaptive choice of nonlinearities.
- Pollard, D. (2001). Chapter 13 from Asymptopia work-in-progress.
- Risk, B. B., Matteson, D. S., Ruppert, D., Eloyan, A., and Caffo, B. S. (2014). An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics*, 70(1):224–236.
- Silva, P. F., Marcal, A. R., and da Silva, R. M. A. (2013). Evaluation of features for leaf discrimination. *Springer Lecture Notes in Computer Science*, Vol. 7950(197-204).
- Sporns, O. (2011). The human connectome: a complex network. *Annals of the New York Academy of Sciences*, 1224(1):109–125.
- Stögbauer, H., Kraskov, A., Astakhov, S. A., and Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Physical Review E*, 70(6):066123.
- Theis, F. J. (2006). Towards a general independent subspace analysis. In *Advances in Neural Information Processing Systems*, pages 1361–1368.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.
- Tohka, J., Foerde, K., Aron, A. R., Tom, S. M., Toga, A. W., and Poldrack, R. A. (2008). Automatic independent component labeling for artifact removal in fMRI. *Neuroimage*, 39(3):1227–1245.
- Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44(10):1–18.